Dissecting Query-Key Interaction in Vision Transformers

Xu Pan[1,2] **Aaron Philip** [3] **Ziqian Xie** [4] **Odelia Schwartz** [1]
[1]University of Miami [2]Harvard University [3]Michigan State University
[4]University of Texas Health Science Center at Houston
xupan@fas.harvard.edu philipaa@msu.edu
ziqian.xie@uth.tmc.edu odelia@cs.miami.edu

Abstract

Self-attention in vision transformers is often thought to perform perceptual grouping where tokens attend to other tokens with similar embeddings, which could correspond to semantically similar features of an object. However, attending to dissimilar tokens can be beneficial by providing contextual information. We propose to analyze the query-key interaction by the singular value decomposition of the interaction matrix (i.e. $\mathbf{W}_q{}^\top \mathbf{W}_k$). We find that in many ViTs, especially those with classification training objectives, early layers attend more to similar tokens, while late layers show increased attention to dissimilar tokens, providing evidence corresponding to perceptual grouping and contextualization, respectively. Many of these interactions between features represented by singular vectors are interpretable and semantic, such as attention between relevant objects, between parts of an object, or between the foreground and background. This offers a novel perspective on interpreting the attention mechanism, which contributes to understanding how transformer models utilize context and salient features when processing images.

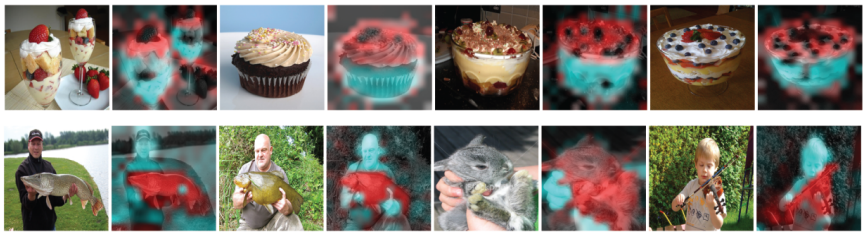

Figure 1: We propose a new way to study query-key interactions via the singular value decomposition of the query-key interaction matrix. Many of the modes (i.e. pairs of singular vectors corresponding to the query and the key respectively), are semantically interpretable. Two example modes are shown. Top row: ViT layer 8 head 7 mode 2. Bottom row: DINO layer 8 head 9 mode 2. The red channel indicates the projection value of embedding onto the left singular vector which corresponds to the query; the cyan channel indicates the projection value of embedding onto the right singular vector which corresponds to the key.

1 Introduction

Vision transformers (ViTs) are a family of models that have significantly advanced the computer vision field in recent years [14]. The core computation of ViTs, self-attention, is designed to promote interactions between tokens corresponding to relevant image features [14]. But this mechanism

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

has different interpretations with open questions such as what "relevant" refers to. Some interpret "relevant" as tokens within the same object. Highlighting objects in attention maps is usually considered a desirable property of ViTs [14, 7, 9]. However, observations in the language domain suggest that self-attention contextualizes tokens, such that the same token has different meanings in different contexts[16]. Contextualization in vision may require a token to receive information not only from same-category tokens, but also from a wider range of different-category tokens such as backgrounds or other objects in the scene. Contextual effects also abound in neuroscience, whereby the responses of neurons and perception are influenced by the context [8, 23, 44, 25, 22, 11, 3, 10]. Therefore, two ideas exist regarding self-attention: a token attends to similar tokens, which could lead to grouping and highlighting the objects; or attends to dissimilar tokens such as backgrounds and different objects, which could lead to stronger contextualization. The former has been supported by many studies, while the latter has been largely ignored in previous studies.

Much like all other deep learning models, though ViTs are successful in many applications, researchers do not have direct access to how information is processed semantically. This issue is particularly important when deploying transformer-based large language models (LLMs) where safety is a priority. As such, there have been studies trying to find feature axes (also known as semantic axes) in the embedding space [18, 5, 6, 13, 34, 19]. A general finding is that embeddings in feedforward layers (i.e. MLP layers) are more semantically interpretable than in self-attention layers [18, 19]. It is believed that the embeddings in the self-attention layers have more superposition, whereas embeddings in the feedforward layers have less superposition due to the expansion of dimensionality [6]. Thus, there has been less focus on finding feature axes in the self-attention layers, and there has been little study addressing interactions between feature axes. In this study, while addressing the role of self-attention, we propose that singular vectors of the query-key interaction are pairs of feature directions. Properties of self-attention heads can be elucidated by studying the properties of their singular modes. We show that those singular vector pairs help semantically explain the interaction between tokens in the self-attention layers.

Our main contributions are as follows:

- We identify a role of self-attention in a variety of ViTs. In many ViTs, especially those with classification training objectives, early layers perform more grouping in which tokens attend more to similar tokens; late layers perform more contextualizing in which tokens attend more to dissimilar tokens. However, this observation has some variability among models and may depend on the training objective: notably, some self-supervised ViTs tend to increase attention to dissimilar tokens in the last few layers.
- We propose a new way to interpret self-attention by analyzing singular modes. Our method goes beyond finding individual feature axes and extends model explainability to the interaction of pairs of feature directions. This approach therefore constitutes enhancing the explainability of transformer models.

In section 2, we state the motivations of this study and list related work. In section 3, we empirically analyze the preference of self-attention between tokens within and between object categories. In section 4, to study the fundamental properties of the query-key interaction, we propose a Singular Value Decomposition method. In section 5, we show that many of the decomposed singular modes are semantic and can be used to interpret the interaction between tokens. In section 6, we discuss the limitations of this study. In section 7, we discuss the main findings and the significance of this study. In the supplementary, we provide an extensive set of visualization examples of the singular modes. Code for this work is available at: https://github.com/schwartz-cnl/DissectingViT.

2 Related work

Attention map properties The properties of attention maps have been studied since the invention of the ViT. The original ViT paper reported that the model attends to image regions that are semantically meaningful, showing that the $[CLS]$ token (i.e. a special token originally designed as the final hidden vector) attends to objects [14]. Later, a study showed that, in a self-supervised ViT named DINO, the $[CLS]$ attention map has a clearer semantic segmentation property, highlighting the object [7]. Following this idea, studies further showed that the attention map of tokens can highlight parts of an object, and subsequently developed a segmentation algorithm by aggregating attention maps

[31, 39]. Another study on the output of self-attention layers indicates that self-attention may perform perceptual grouping of similar visual objects, rather than highlighting a salient singleton object that stands out from other objects in the image [29]. Most of these studies focus on the $[CLS]$ token attention map or on the outputs of attention maps. Our study, in contrast, seeks to interpret the interactions between tokens within the self-attention layers, to gain insights about properties such as grouping and contextualization.

Contextualization Our study is inspired by contextual effects in visual neuroscience, in which neural responses are modulated by the surrounding context [3, 8, 44]. For instance, the response of a cortical visual neuron in a given location of the image is suppressed when the surrounding inputs are inferred statistically similar, but not when the surround is inferred statistically different, thereby highlighting salient stimuli in which the center stands out from the surround [25, 12]. Some of these biological surround contextual effects have been observed in convolutional neural networks [28, 32]. Here our goal is not to address biological neural contextual effects in ViTs, but to dissect contextual interactions in the self-attention layers. It is known that language transformer models have a strong ability to contextualize tokens [16]. However, it's not clear what kinds of contextualization emerge in the ViT. In this study, we seek to understand what kinds of interactions occur between a token and other tokens that carry important contextual information, possibly representing different objects, different parts of an object, or the background.

Finding feature axes Finding feature axes is crucial for understanding and controlling model behavior. Since a study found semanticity in the embeddings of feedforward layers in LLMs [18], studies have primarily focused on identifying feature axes in the feedforward layers, and to a lesser extent, in self-attention layers. Similar to the findings in LLMs, a ViT study found that feedforward layers have less mixed concepts and can generate interpretable feature visualizations [19]. Bills et al. proposed a gradient-based optimization method to find explainable directions in LLMs [5]. Later, Bricken et al. proposed a simpler method of sparse autoencoder [6]; though see [21]. These methods have not been extensively applied to ViT studies.

Some studies focused on finding feature directions in the ViTs' self-attention layers. In downstream tasks such as semantic segmentation, researchers empirically found that choosing the key embeddings as features leads to the best performance [37, 2, 1]. A study proposed that the singular value decomposition of the weight matrix is a natural way to find feature directions in any neural network [34]. But they only focused on single feature directions (right singular vectors), and did not consider the feature interaction in the context of self-attention. Another study suggested that singular vectors of value weights and feedforward weights can be used as features in LLMs, but they did not analyze the query-key interaction matrix [30]. Another study in the language domain proposed a singular vector decomposition on the union of the query and key embeddings, but not on the query and key weights [27].

There has been limited work going beyond single features to studying query-key interactions. A study focusing on LLMs proposed that the corresponding columns of query and key matrices are interpretable as pairs [13]. However, this approach does not find features beyond the canonical basis of the query and key embeddings. Another study inferred query-key interactions by jointly visualizing them in a low dimensional space, but their method does not find interacting feature axes [42]. Here, in contrast to previous works, we utilize the singular value decomposition to study the query-key interactions. We propose that left and right singular vectors of the query-key interaction matrix can be seen as pairs of interacting feature directions, and study their properties in ViTs.

3 Grouping or contextualizing

Firstly, we empirically study whether an image token (i.e. a patch in the image) attends to tokens belonging to the same objects, different objects, or background. We utilized a dataset that has been applied to studying visual salience [24], namely the Odd-One-Out (O3) dataset [29]. This dataset was also used by Mehrami et al. [29] in their study but they only focused on the output of the attention layers. However, we use a different experimental design that focuses on the attention maps of image tokens. The dataset consists of 2001 images that have a group of similar objects (distractors) and a distinct singleton object (target) (Fig 2 A). Our goal is to examine if the attention map of a

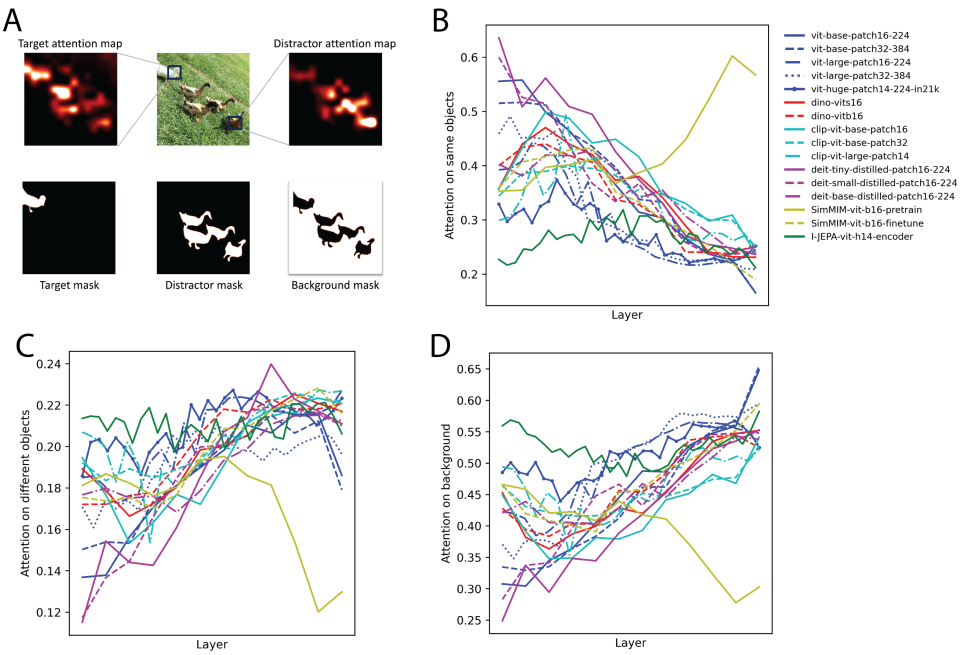

Figure 2: Attention preference in the Odd-One-Out (O3) dataset [24]. A. An example from the O3 dataset. Two tokens are chosen to correspond to the target and distractor in the image. Attention maps using two tokens as queries are computed. We examine the overlap between the attention map of the target, and each of the mask labels of the target, distractor, and background masks. Similarly, we examine the overlap between the attention map of the distractor, and each of the mask labels of the distractor, target, and background. B. Ratio of attention on the same objects (target-target and distractor-distractor attention). The x-axis is normalized layer numbers, from early layers (left) to late layers (right). C. Ratio of attention on the different objects (target-distractor and distractor-target attention). D. Ratio of attention on the background (target-to-background and distractor-background attention)

token of one category (target or distractors) covers more of the same category, different category, or background.

We chose to study 16 different ViT models from 6 families: the original ViT [14], DeiT which uses distillation to learn from a teacher model [38], CLIP which is jointly trained with a text encoder [33], and DINO [7], SimMIM [40], I-JEPA [4] which are self-supervised models with either contrastive or mask prediction loss.

In this study, the "attention score" is defined as the dot product of every query and key pair, which has the shape of the number of tokens by the number of tokens and is defined per attention head. The "attention map" is the softmax of each query's attention score reshaped into a 2D image, which is defined per attention head and token. For each image in the dataset, two tokens are chosen to represent the target and distractor. They are at the location of the maximum value of the down-scaled target or distractor mask. Two attention maps are obtained using the two tokens, each is normalized to sum to 1. Inner products are computed between the two attention maps and three masks, which can be interpreted as the ratio of attention of an object (target or distractor) on the same object, different object, or background. We use target-target, target-distractor, target-background, distractor-target, distractor-distractor, and distractor-background attention to denote the 6 inner products. This measure is computed per layer, head, and image. The averaged measure is shown in Fig 2. Target-target and distractor-distractor attention are categorized as "attention on same objects"; target-distractor and distractor-target attention are categorized as "attention on different objects"; target-to-background and distractor-to-background attention are categorized as "attention on background". The attention on the same objects should be dominant if attention is to perform grouping. We find a trend that in most ViTs the attention on the same objects is dominant in early layers; while there is a trend that in

the deeper layers attention gradually increases on the contextual features such as the background or different objects. However, this observation has some variability among models and may depend on the training objective. For example, the self-supervised SimMIM pre-trained on pixel-level mask prediction shows increased attention on the same objects in later layers. Interestingly, this trend disappears after fine-tuning on a classification task.

This result provides new evidence that self-attention considers contextual features as much or more than similar features in deeper layers. In most ViTs, especially with those with classification training objectives, self-attention prefers the same objects in early layers; in deeper layers, self-attention shifts to contextual information. As far as the authors are aware, this finding has not been reported in previous ViT studies [14, 7, 39, 29].

4 Singular value decomposition of query-key interaction

4.1 Formulation

In the previous section, we empirically study the allocation of self-attention and find that self-attention does not only do grouping. In this section, we try to find whether this self-attention property can be better understood by analyzing the underlying computation. The self-attention computation is formulated as below, following the convention in the field. Each token is first transformed into three embeddings, namely query, key and value. The output of a self-attention layer is the sum of values weighted by some similarity measures between query and key. The original transformer model used the softmax of the dot-product of the key and query [14]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}^\top \mathbf{K}}{\sqrt{d_k}})\mathbf{V}$$

where $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ denote the query, key, and value embeddings. They are calculated from linearly transforming the input sequence $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_L\} \in \mathbb{R}^{d \times L}$, where d is the input embedding size, L is the sequence length,

$$\mathbf{Q} = \mathbf{W}_q \mathbf{X} \in \mathbb{R}^{d_k \times L}$$
$$\mathbf{K} = \mathbf{W}_k \mathbf{X} \in \mathbb{R}^{d_k \times L}$$
$$\mathbf{V} = \mathbf{W}_v \mathbf{X} \in \mathbb{R}^{d_v \times L}$$

where $\mathbf{W}_q \in \mathbb{R}^{d_k \times d}$, $\mathbf{W}_k \in \mathbb{R}^{d_k \times d}$, $\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$ are trainable linear transformations that transform the input embedding to the key, query, and value space. Sometimes a bias term is also added to the transformation. Since the bias term does not depend on the input embedding, we do not include it in our analysis of token interactions. In the formula of the attention output, the part that contains the query and key interaction is named the attention score. In this case which is based on the dot-product, the attention score between two tokens $\mathbf{x}_i$ (query) and $\mathbf{x}_j$ (key) is

$$a_{ij} = \mathbf{q}_i^\top \mathbf{k}_j = \mathbf{x}_i^\top {\mathbf{W}_q}^\top \mathbf{W}_k \mathbf{x}_j$$

The attention score solely depends on the combined matrix ${\mathbf{W}_q}^\top \mathbf{W}_k$ as a whole [15], which represents the query-key interaction. To better understand the behavior of this bilinear form, we factor the matrix using the singular value decomposition,

$$\mathbf{W}_q^\top \mathbf{W}_k = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where $\mathbf{U} = \{\mathbf{u}_1, ..., \mathbf{u}_{d_k}\} \in \mathbb{R}^{d \times d_k}$ is the left singular matrix composed of left singular vectors, $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_{d_k}\} \in \mathbb{R}^{d \times d_k}$ is the right singular matrix composed of right singular vectors, $\mathbf{\Sigma} = \text{diag}(\sigma_1, ..., \sigma_{d_k}) \in \mathbb{R}^{d_k \times d_k}$ is a diagonal matrix composed of singular values. We will refer to *the nth singular mode* as the set $\{\mathbf{u}_n, \sigma_n, \mathbf{v}_n\}$. Then the attention score between two tokens can be decomposed into singular modes.

$$\mathbf{x}_i^\top {\mathbf{W}_q}^\top \mathbf{W}_k \mathbf{x}_j = \sum_{n=1}^{d_k} \mathbf{x}_i^\top \mathbf{u}_n \sigma_n \mathbf{v}_n^\top \mathbf{x}_j$$

Consider the input embeddings projected onto the left and right singular vectors, i.e. $\mathbf{x}^\top \mathbf{u}_n$ and $\mathbf{x}^\top \mathbf{v}_n$. The attention score is non-zero when the two embeddings have a non-zero dot-product with the

corresponding left and right singular vectors within the same singular mode. In other words, if one embedding happens to be in the direction of a left singular vector, it only attends to tokens that have a component of the corresponding right singular vector. It can be thought of as a left singular vector "query" looking for its right singular vector "key".

4.2 Similarity between left and right singular vectors

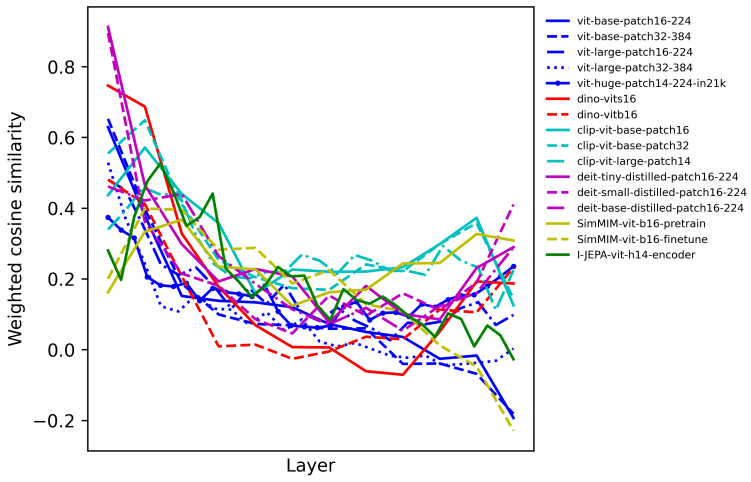

Figure 3: Cosine similarity between left and right singular vectors. The cosine similarity is computed per head and singular mode. The weighted average value of cosine similarity is computed with weights of corresponding singular values.

To determine if self-attention performs grouping or combines contextual information, we examine whether tokens in different layers have higher attention scores with similar tokens or dissimilar tokens. This can be measured for each singular mode by how much the left singular vector is aligned with the right singular vector, more specifically, the cosine similarity between the left singular vector and the right singular vector. A high cosine similarity value means tokens attend to similar tokens (to itself if the value is 1); a low value means tokens attend to dissimilar tokens (to orthogonal tokens if 0; to opposite tokens if negative). The average cosine similarity is weighted by the singular values with the assumption that singular modes with higher singular values are more influential to the total attention score $\cos_{avg} = \sum_i \frac{\sigma_i}{\sum \sigma_j} \cos_i$. We find that the averaged cosine similarity is high in early layers, and there is a decreasing trend in deeper layers (Fig 3). In some models, the averaged cosine similarity drops to 0 in some middle layers. The cosine similarity distribution and singular value spectrum of the vit-base-patch16-224 model is provided in the Supplementary Figures S1 and S2.

Though we find a general trend that attention changes from attending more to the similar tokens to dissimilar tokens from early layers to late layers, some ViTs have a more complex trend that increases attention to similar tokens in the last few layers (Fig 3). Models that have this “concave” trend are SimMIM-vit-b16-pretrain, Dino models, Deit models, and huge ViT models. Most of them either have self-supervised objectives or distillation regularizations. We hypothesize that the last layers may behave differently because they are closer to the training target, and so the training objective may have more influence. We think that self-supervised objectives, such as reconstructing masked patches, require stronger consistency between tokens, and thus more attention is allocated to similar tokens in the higher layers; while the classification objective requires gathering information from different aspects of a scene, and thus more attention is allocated to dissimilar tokens. This hypothesis is supported by the cosine similarity plot (Fig 3) of the SimMIM models, which shows in the last few layers of the pre-trained model increased attention to similar features. This matches the observation in the literature, that the SimMIM model has more local attention [41]. However, we find that the SimMIM model fine-tuned on ImageNet classification has a trend of decreased attention to similar features, similar to most of the classification models. Although I-JEPA is trained with a self-supervised objective predicting latent representations, the cosine similarity for the I-JEPA

encoder does not show increased attention to similar tokens in the last few layers. The I-JEPA model is known to have excellent linear-probing performance, and thus we think it may behave more similarly to a classification model. The self-supervised objective of I-JEPA may be more apparent in the I-JEPA predictor (also a transformer). When we run the cosine similarity analysis on the predictor module instead of the encoder, we find that the cosine similarity is overall high (Supplementary FigureS3). The role of the training objective on internal model behavior is an interesting topic for future research.

It is known that embeddings in transformer models are to some extent anisotropic [17, 26, 20], which means the expected value of cosine similarity of two random sampled inputs tends to be positive. We indeed find anisotropy effects in all the models we examined using cosine similarity (Supplementary Figure S4) (though see other metrics [35]). If we treat anisotropy level as a baseline for cosine similarity, the effect shown in Fig 3 still exists but the self-attention is less biased to similar tokens (Supplementary Figure S4).

There is a further implication of the singular value decomposition approach. The left and right singular vectors of each attention head are two incomplete orthonormal bases of embedding. We suggest that these bases are feature directions since they are intrinsic properties of the self-attention layer. The query and key embeddings can be made arbitrary, since one can change the basis without affecting the attention score. However, the singular vectors are invariant to the change of basis. If an invertible matrix $\mathbf{A} \in \mathbb{R}^{d_k \times d_k}$ acts on the query and key weights as $\mathbf{W}_q \to \mathbf{A}^\top \mathbf{W}_q$ and $\mathbf{W}_k \to \mathbf{A}^{-1}\mathbf{W}_k$, then the attention score does not change but the query and key embeddings change. The singular vector decomposition of $(\mathbf{A}^\top \mathbf{W}_q)^\top \mathbf{A}^{-1}\mathbf{W}_k$ stays the same as decomposing ${\mathbf{W}_q}^\top \mathbf{W}_k$. Thus singular vectors are uniquely special and may show interesting properties. Due to the sign ambiguity of the singular value decomposition, we consider the opposite directions of singular vectors also as feature directions.

5 Semanticity of singular modes

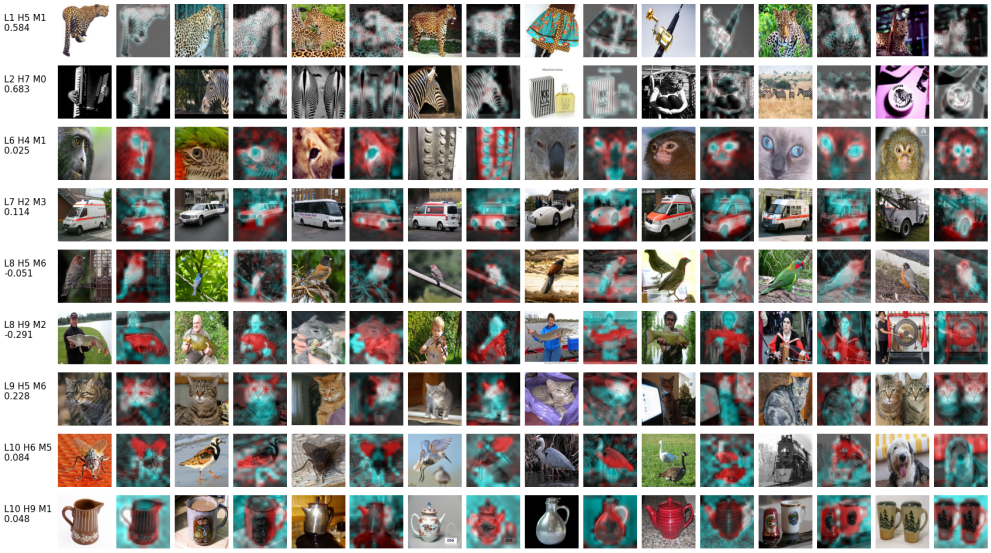

Figure 4: Examples of optimal attention images of singular modes and query and key map in dino-vitb16. Optimal attention images are found from the Imagenet validation set that induce the largest attention score (sorted by the product of the maximum of query map and maximum of key map). The red and cyan (i.e. green and blue) channels are the projection values of embedding onto the left and right singular vectors of a singular mode. They correspond to query and key. The white area is where the query map and key map overlap. The name code we assign to singular modes specifies the layer, head, and mode numbers. For example, "L1 H5 M1" means layer 1, head 5, and mode 1. The value below indicates the cosine similarity between the left and right singular vectors.

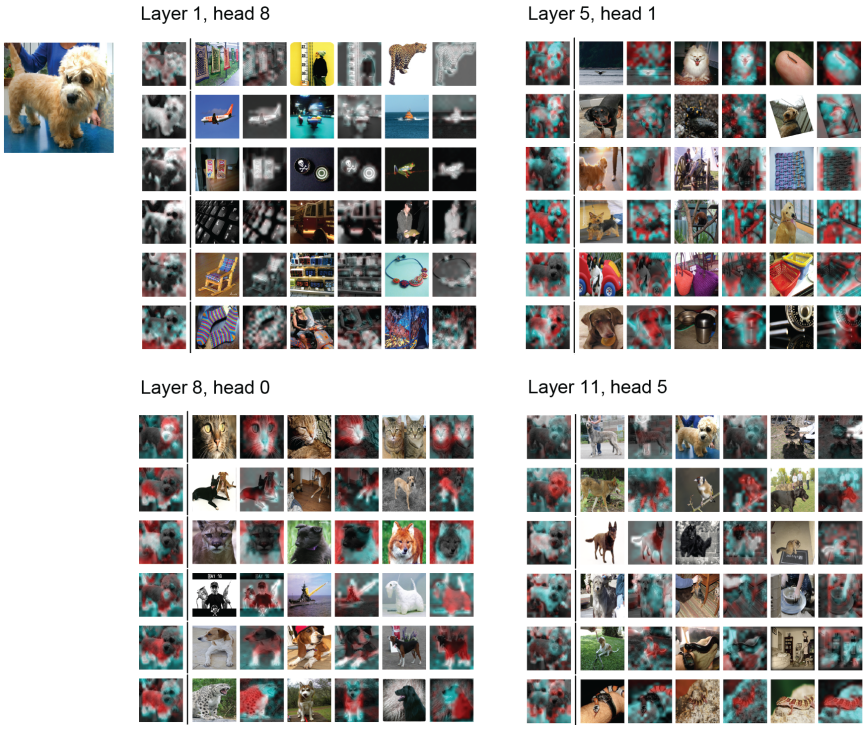

Figure 5: Visualization of a single image with multiple modes. We pick an example dog image from the ImageNet dataset and use the dino-vitb16 model. Top 6 modes (ordered by the contribution to the attention score) for example layers and heads are shown. See Supplementary Figure S17 for extended mode visualizations of this image.

The singular value decomposition of self-attention offers an intuitive way to explain the self-attention layer. A feature represented by a left singular vector attends to the feature represented by the corresponding right singular vector. The feature of a singular vector can be found by finding the image that has the maximum embedding projection on the singular vector. Similarly, the typical interactions of a singular mode can be identified by finding the image that has the maximum product of the projections on a singular vector pair. Previous studies on the explainability of deep learning models only focused on the explainability of single neurons or individual feature axes. The singular value decomposition extends model explainability to the interaction of pairs of "neurons" (i.e. singular vectors). Note that this is very different from the standard approach of visualizing the attention map of the $[CLS]$ token without addressing interactions between tokens [14, 7, 31].

Some example modes from dino-vitb16 are shown in Fig. 4. For each mode, we show the top 8 images in the Imagenet (Hugging Face version) [36] validation set that induce the largest attention score. For each image, a query map (red channel in the figure) and a key map (cyan channel in the figure) are obtained by projecting the embedding onto the left and right singular vectors. Each map tells what information the left or right singular vector represents. Jointly, the highlighted regions in the query map attend to the highlighted regions in the key map. In other words, the information in the highlighted regions of the key map flows to the highlighted regions of the query map. More examples are shown for a range of ViT architectures in the Supplementary Figures S5 - S16.

In early layers, singular vectors usually represent low-level visual features like color or texture, and sometimes positional encoding. In higher layers, singular vectors can represent more complex visual features like parts of objects or whole objects. As shown in the previous sections, high attention scores can be induced between similar tokens (more often in early layers) or dissimilar tokens (more often in late layers). The correspondence to image structure for similar and dissimilar tokens can be seen in the query and key maps. For the modes with high cosine similarity, query and key maps are similar which could represent color, texture, parts, objects, or positional encoding. For the modes

with low cosine similarity, query and key maps look different which could represent different object parts, different objects, or foreground and background. Some examples include: in "L6 H4 M1" the animal face (query) attends to eyes, nose and mouth (key); in "L7 H2 M3" the lower part of a car attends to the upper part of a car and wheels; in "L8 H9 M2" the fish or other things in hand attend to human; in "L10 H9 M1" the kettle attends to its background.

To show the hierarchical information process across layers, we show an example dog image and example attention heads along with optimal images for top modes in Fig. 5. The late layers capture more semantic information such as the parts of a dog or animal, and a hand with a dog. The early layers capture low-level properties like color. We show more examples in Supplementary Figure S17-S19.

The attention between dissimilar tokens could be thought of as providing contextual information to a given token. In the part-to-part case, finding more parts of an object increases the confidence of finding the object and helps merge smaller concepts into a larger concept. In the object-to-object case, an object attending to a different object could add additional attributes to it, for example, a fish attending a human may add the attribute "be held" to the fish tokens, which helps understanding of the whole scene. These interactions between tokens, though conceptually simple, as far as the authors are aware, have not been reported before this study. This result further supports the idea that self-attention combines contextual information from dissimilar tokens such as backgrounds or different objects.

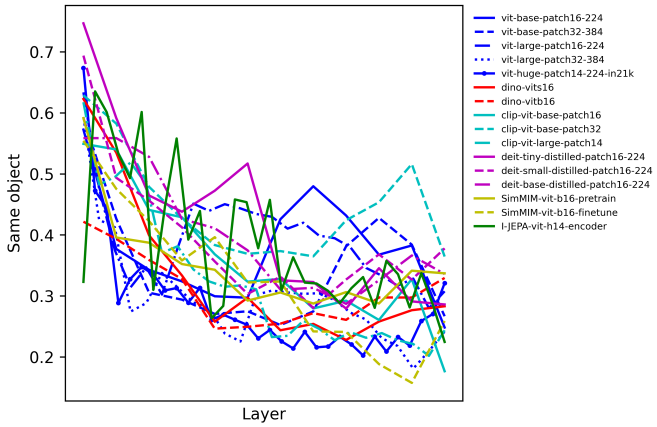

Figure 6: The probability that the left and right singular vectors highlight the same object in maximum attention images.

Finally, we study whether tokens prefer to attend to the same object or different objects at the singular mode level. We choose to use a semantic segmentation dataset, namely ADE20K [43]. We first find the top 5 images that induce maximum attention of a singular mode, then find the optimal objects in each image that have the maximum projections on the left and right singular vectors per object area. The probability of the left and right singular vectors having the same optimal object is computed with the weight of singular values, following the same method in the previous experiment. We find that, in early layers, there is a higher probability that the left and right singular vectors attend to the same object; in late layers, the probability is lower, though the variability between models is considerably large (Fig. 6). This result further supports that self-attention performs more grouping in early layers; in late layers, tokens attend to different objects which could contextualize the token with background information.

6 Limitation

We are aware of some limitations of this study and interesting open questions that remain. There is behavioral variability between the models, which may be due to the distinct training objectives. Identifying how the training paradigm alters the learned embedding space is a potential future

direction to explore. We have focused on the query-key interactions in the self-attention, and future studies could address the role of the value matrix.

7 Discussion

Inspired by the observation that self-attention gathers information from relevant tokens within an object, and the importance of contextualization in neuroscience, we study fundamental properties of token interaction inside self-attention layers in ViTs. Both empirical analysis of the Odd-One-Out (O3) dataset, and singular decomposition analysis of singular modes for the Imagenet dataset, show that in early layers the attention score is higher between similar tokens, while in late layers the attention score is higher between dissimilar tokens.

The singular decomposition analysis provides a new perspective on the explainability of ViTs. Two directions (left and right singular vectors) in the embedding space could be analyzed in pairs to interpret the interaction between tokens. Using this method, we find interesting semantic interactions such as part-to-part attention, object-to-object attention, and foreground-to-background attention which have not been reported in previous studies. Our reported findings provide evidence that self-attention in vision transformers is not only about gathering information between tokens with similar embeddings, but a variety of interactions between a token and its context. The method of analyzing singular vectors can be easily adapted to study token interactions in transformer networks trained on other modalities like language. Adapting this method to real-world applications can increase transparency of what the transformer models are capturing.

Acknowledgements

A.P. was supported by the Research Experiences for Undergraduates (REU) Site Scientific Computing for Structure in Big or Complex Datasets, NSF grant CNS-1949972. O.S. was funded by the University of Miami Provost Research Award.

References

[1] H. Adeli, S. Ahn, N. Kriegeskorte, and G. Zelinsky. Affinity-based attention in self-supervised transformers predicts dynamics of object grouping in humans. *arXiv preprint arXiv:2306.00294*, 2023.

[2] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[3] A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff. Circuits and mechanisms for surround modulation in visual cortex. *Annual review of neuroscience*, 40: 425–451, 2017.

[4] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

[5] S. Bills, N. Cammarata, D. Mossing, H. Tillman, L. Gao, G. Goh, I. Sutskever, J. Leike, J. Wu, and W. Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2023.

[6] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2, 2023.

[7] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[8] J. R. Cavanaugh, W. Bair, and J. A. Movshon. Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of neurophysiology*, 88(5): 2547–2556, 2002.

[9] X. Chen, C.-J. Hsieh, and B. Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. In *International Conference on Learning Representation*, 2022.

[10] O.-H. Choung, A. Bornet, A. Doerig, and M. H. Herzog. Dissecting (un) crowding. *Journal of vision*, 21(10):10–10, 2021.

[11] C. W. Clifford and G. Rhodes. *Fitting the mind to the world: Adaptation and after-effects in high-level vision*, volume 2. Oxford University Press, 2005.

[12] R. Coen-Cagli, A. Kohn, and O. Schwartz. Flexible gating of contextual influences in natural vision. *Nature neuroscience*, 18(11):1648–1655, 2015.

[13] G. Dar, M. Geva, A. Gupta, and J. Berant. Analyzing transformers in embedding space. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL https://aclanthology.org/2023.acl-long.893.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[15] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[16] K. Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006.

[17] K. Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[18] M. Geva, R. Schuster, J. Berant, and O. Levy. Transformer feed-forward layers are key-value memories. In M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446.

[19] A. Ghiasi, H. Kazemi, E. Borgnia, S. Reich, M. Shu, M. Goldblum, A. G. Wilson, and T. Goldstein. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*, 2022.

[20] N. Godey, É. de la Clergerie, and B. Sagot. Is anisotropy inherent to transformers? *arXiv preprint arXiv:2306.07656*, 2023.

[21] R. Huben. Research Report: Sparse Autoencoders find only 9/180 board state features in OthelloGPT — aizi.substack.com. https://aizi.substack.com/p/research-report-sparse-autoencoders. [Accessed 24-03-2024].

[22] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001.

[23] H. Jones, W. Wang, and A. Sillito. Spatial organization and magnitude of orientation contrast interactions in primate v1. *Journal of neurophysiology*, 88(5):2796–2808, 2002.

[24] I. Kotseruba, C. Wloka, A. Rasouli, and J. K. Tsotsos. Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations. In *British Machine Vision Conference (BMVC)*, 2019.

[25] Z. Li. Contextual influences in v1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences*, 96(18):10530–10535, 1999.

[26] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[27] T. Lieberum, M. Rahtz, J. Kramár, G. Irving, R. Shah, and V. Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

[28] T. Marques, M. Schrimpf, and J. J. DiCarlo. Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *bioRxiv*, pages 2021–03, 2021.

[29] P. Mehrani and J. K. Tsotsos. Self-attention in vision transformers performs perceptual grouping, not attention. *arXiv preprint arXiv:2303.01542*, 2023.

[30] B. Millidge and S. Black. The singular value decompositions of transformer weight matrices are highly interpretable. In *AI Alignment Forum*, 2022.

[31] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[32] X. Pan, A. DeForge, and O. Schwartz. Generalizing biological surround suppression based on center surround similarity via deep neural network models. *PLoS Computational Biology*, 19 (9):e1011486, 2023.

[33] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[34] A. Radhakrishnan, D. Beaglehole, P. Pandit, and M. Belkin. Mechanism for feature learning in neural networks and backpropagation-free machine learning models. *Science*, 2024.

[35] W. Rudman, N. Gillman, T. Rayne, and C. Eickhoff. Isoscore: Measuring the uniformity of embedding space utilization. *arXiv preprint arXiv:2108.07344*, 2021.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[37] O. Siméoni, G. Puy, H. V. Vo, S. Roburin, S. Gidaris, A. Bursuc, P. Pérez, R. Marlet, and J. Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[39] Y. Wang, X. Shen, S. X. Hu, Y. Yuan, J. L. Crowley, and D. Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[40] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[41] Z. Xie, Z. Geng, J. Hu, Z. Zhang, H. Hu, and Y. Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14475–14485, 2023.

[42] C. Yeh, Y. Chen, A. Wu, C. Chen, F. Viégas, and M. Wattenberg. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[43] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[44] C. M. Ziemba, J. Freeman, E. P. Simoncelli, and J. A. Movshon. Contextual modulation of sensitivity to naturalistic image structure in macaque v2. *Journal of neurophysiology*, 120(2): 409–420, 2018.

A Supplemental material

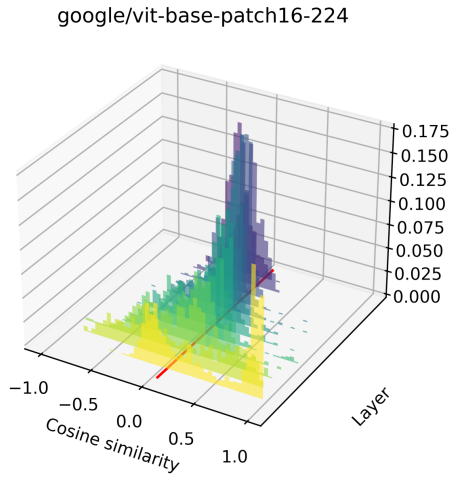

Figure S1: Histogram of cosine similarity between the left and right singular vector in ViT-base-patch16-224. The yellow layers are earlier layers; the blue layers are later layers. The red line indicates 95% confidence interval, which is calculated from embeddings sampled from a random distribution.

Figure S2: Singular value spectrum (blue) and cosine similarity (red) in ViT-base-patch16-224. Row number indicates layer number. Column number indicates head number. The dotted line indicates 95% confidence interval, which is calculated from embeddings sampled from a random distribution.

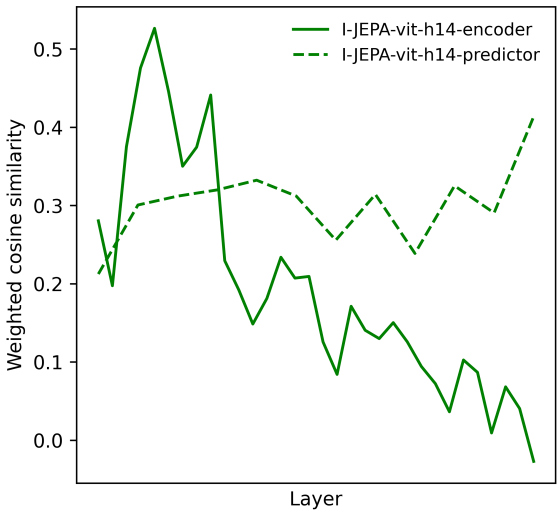

Figure S3: Cosine similarity between left and right singular vectors of the I-JEPA encoder and predictor modules.

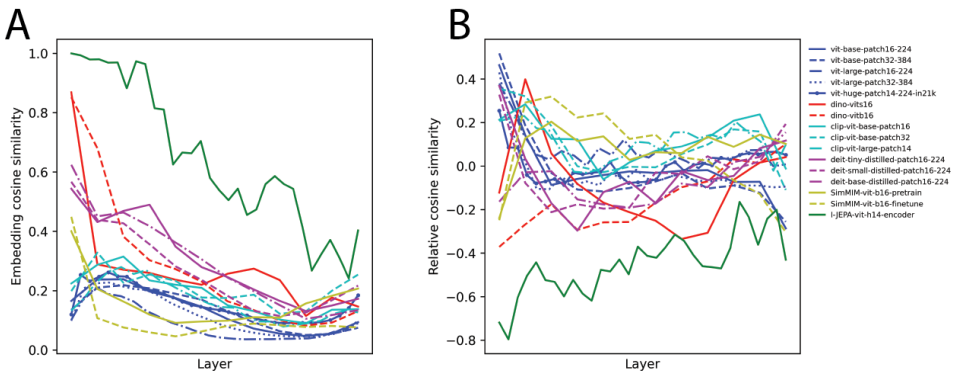

Figure S4: Anisotropy effects in ViTs. A. Averaged embedding cosine similarity between the center tokens of different images from the Imagenet validation set. Consisting with previous studies, the cosine similarities are all positive, which is referred to as anisotropy or cone effect. B. Considering A as the baseline, relative cosine similarity is defined as subtracting cosine similarity between left and right singular vectors by the embedding cosine similarity in A.

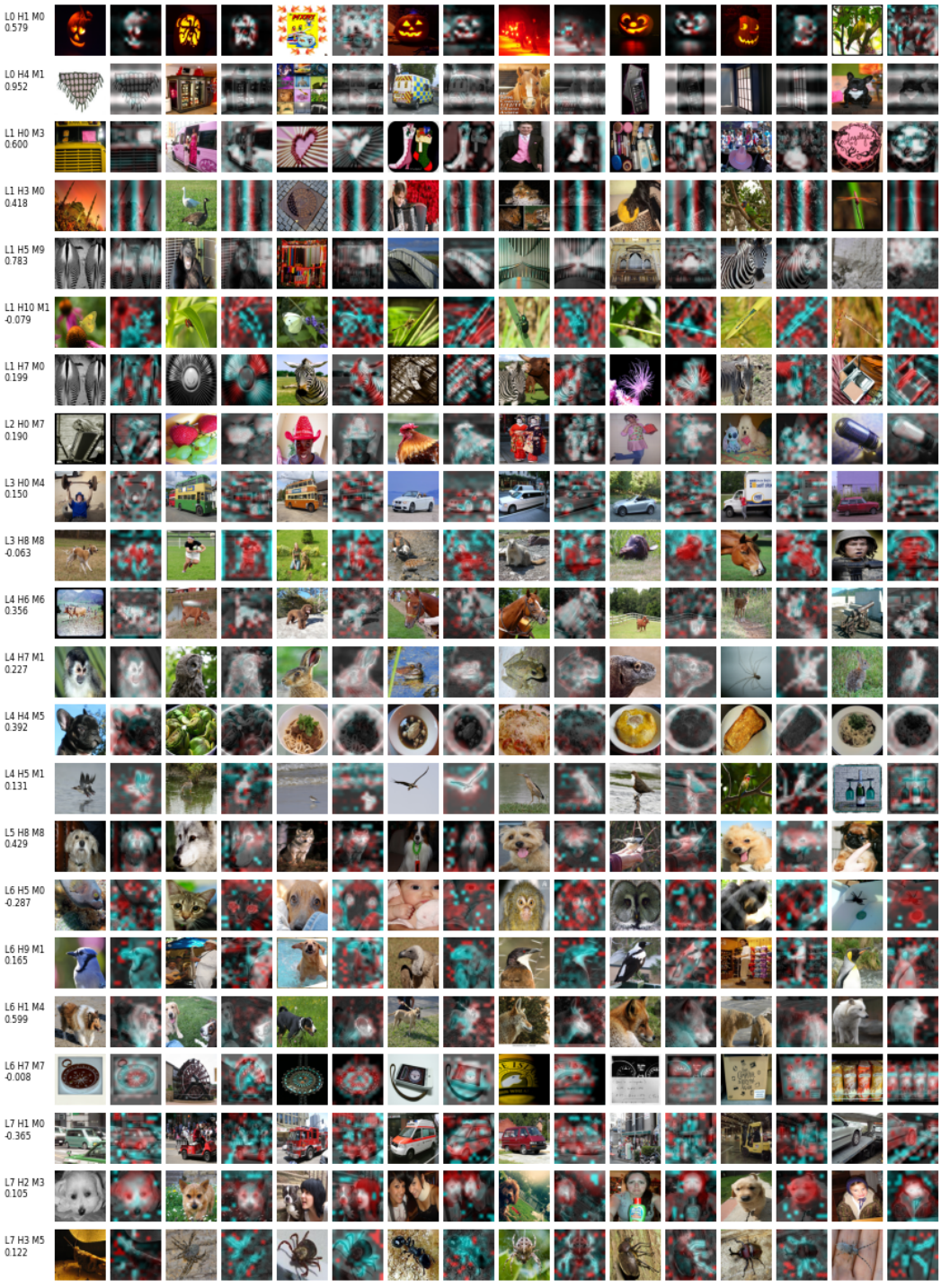

Figure S5: Examples of semantic singular modes in ViT-base-patch16-224 (part 1).

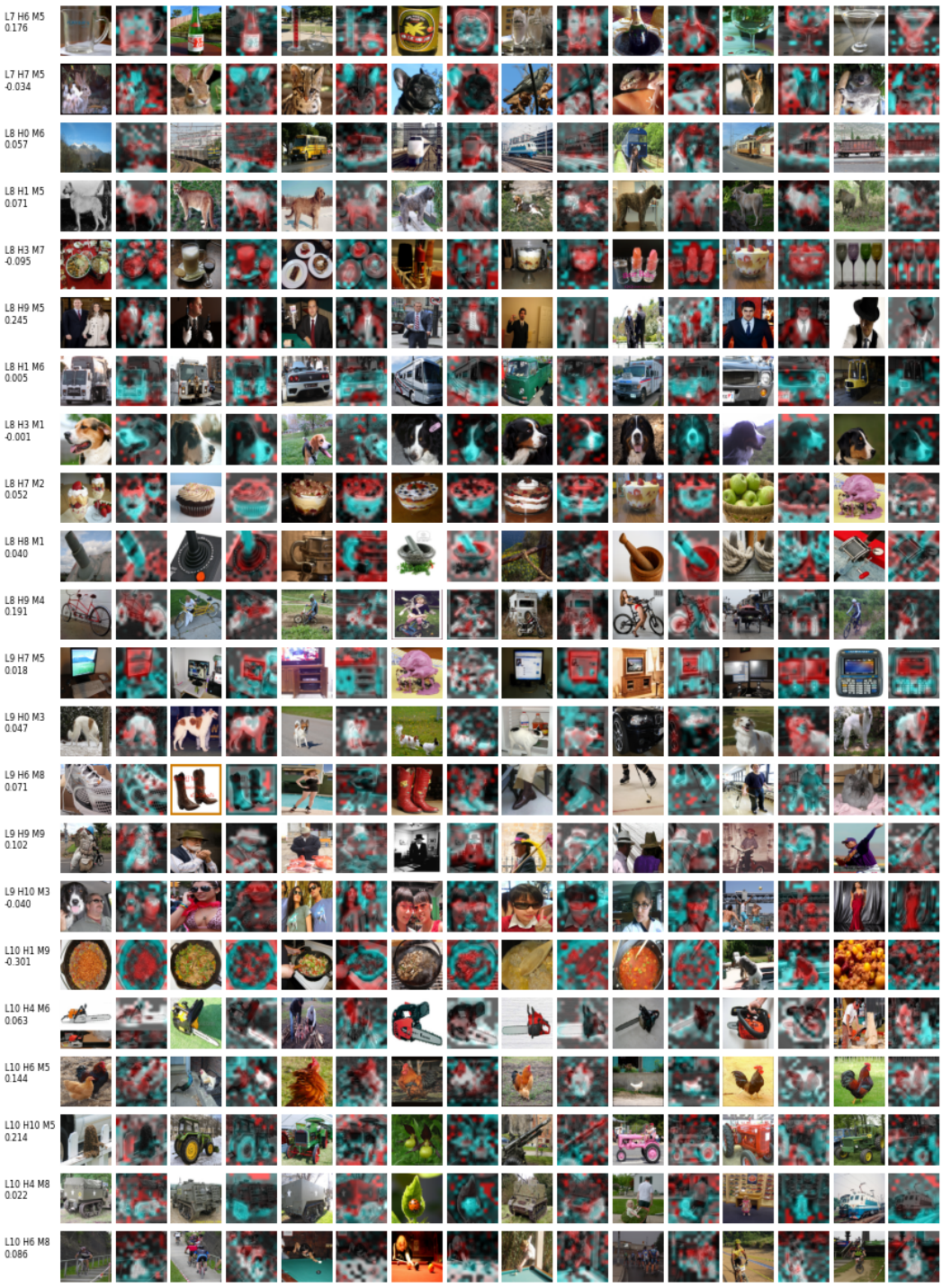

Figure S6: Examples of semantic singular modes in ViT-base-patch16-224 (part 2).

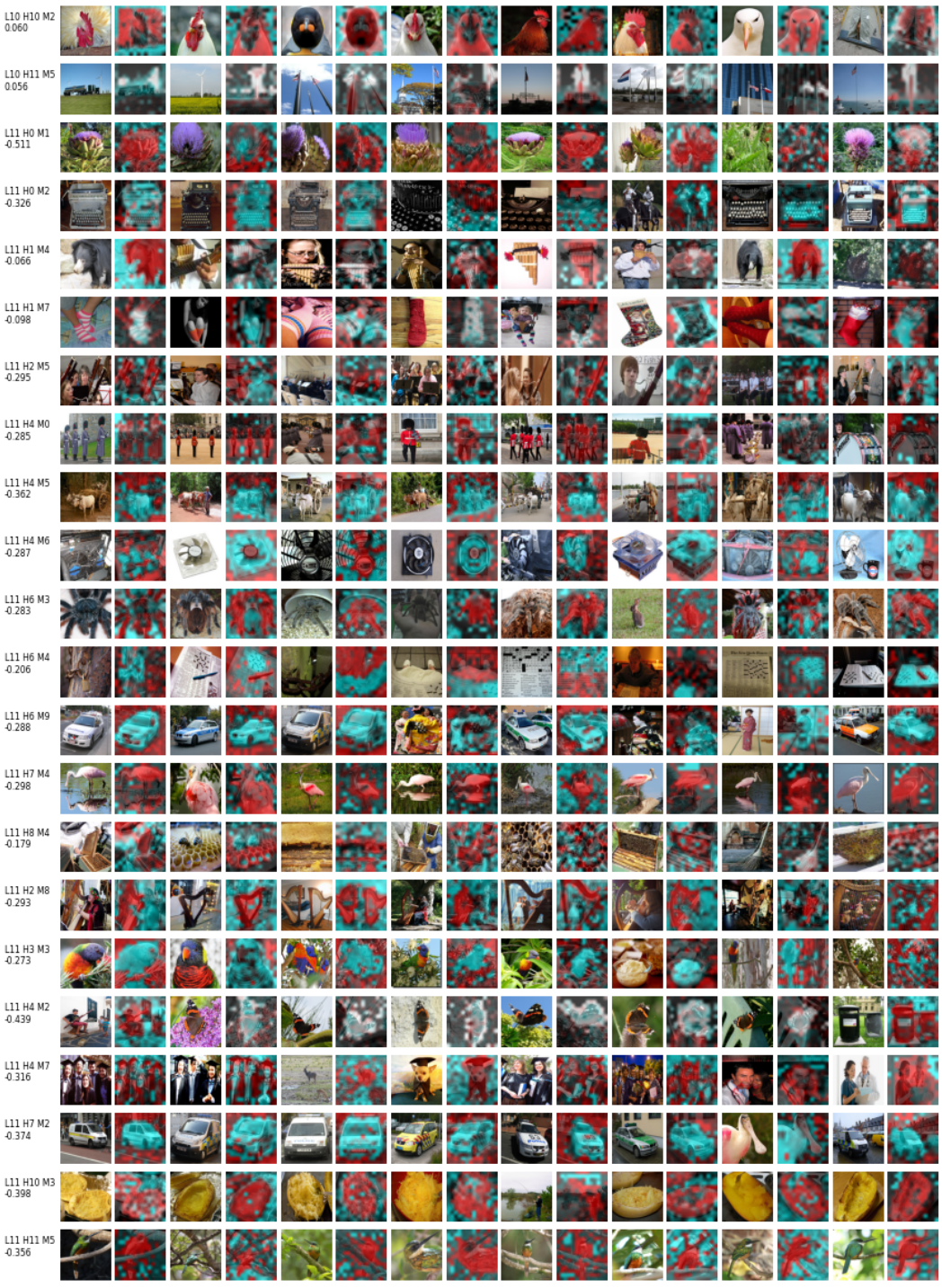

Figure S7: Examples of semantic singular modes in ViT-base-patch16-224 (part 3).

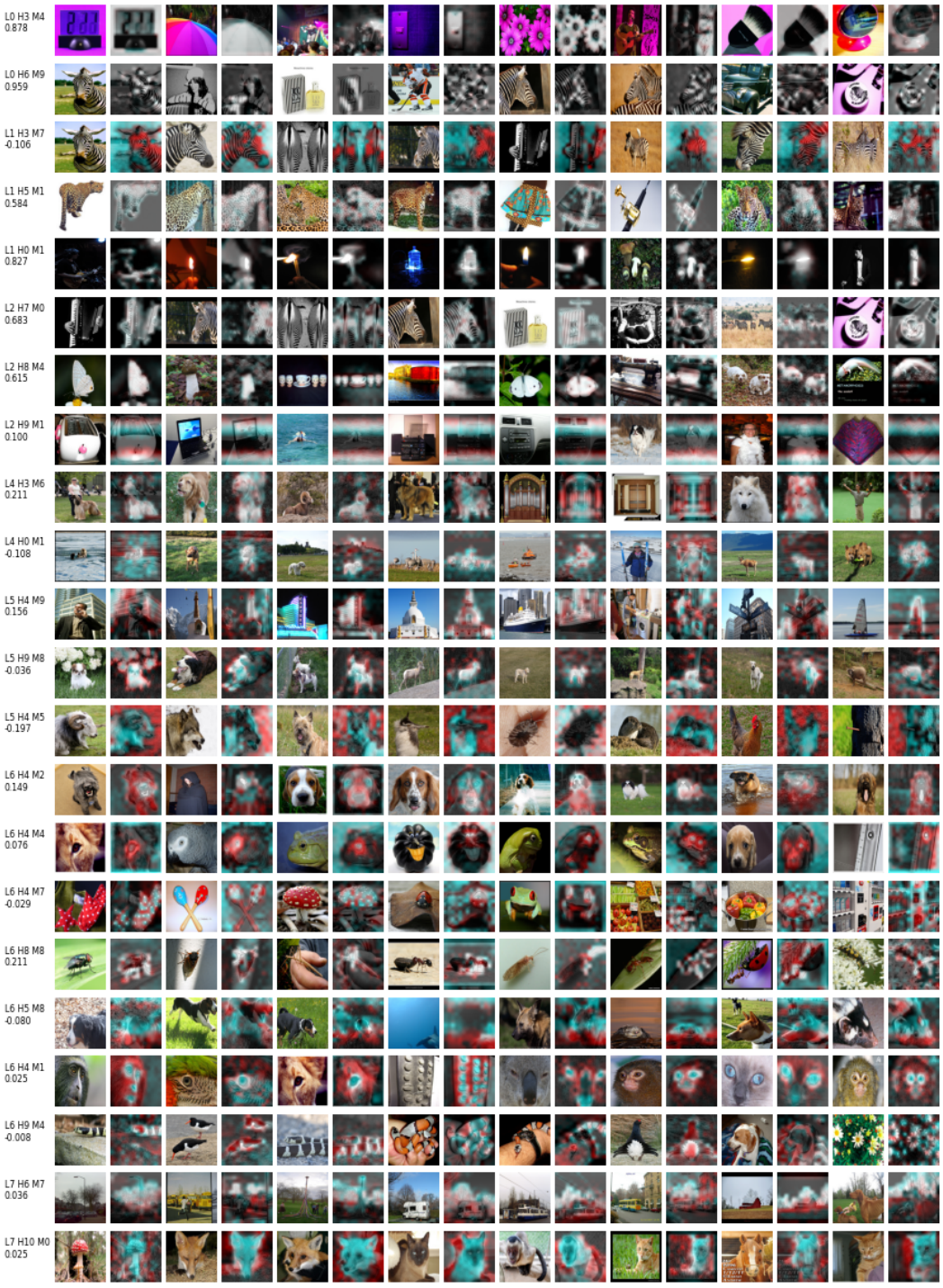

Figure S8: Examples of semantic singular modes in dino-vitb16 (part 1).

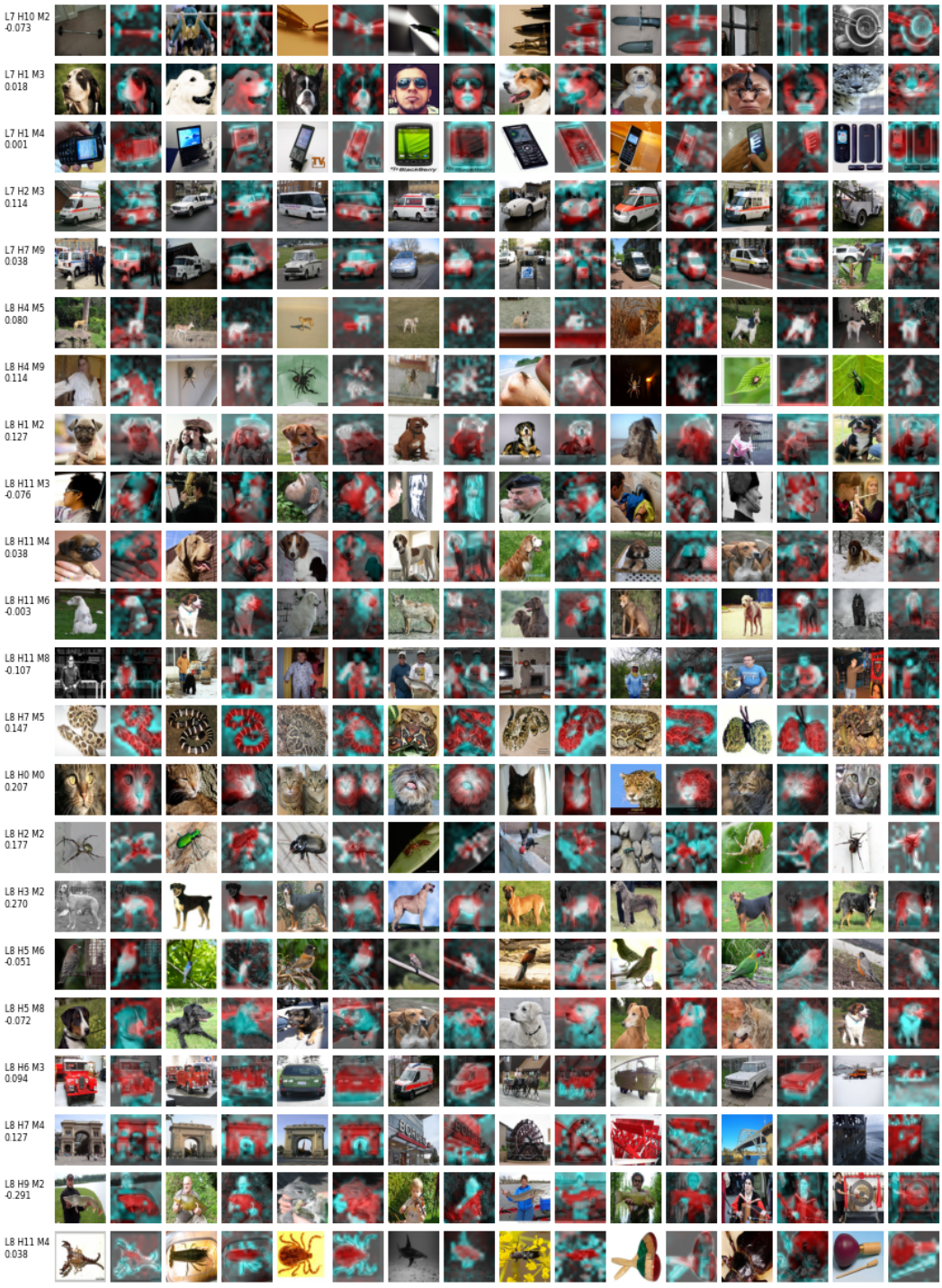

Figure S9: Examples of semantic singular modes in dino-vitb16 (part 2).

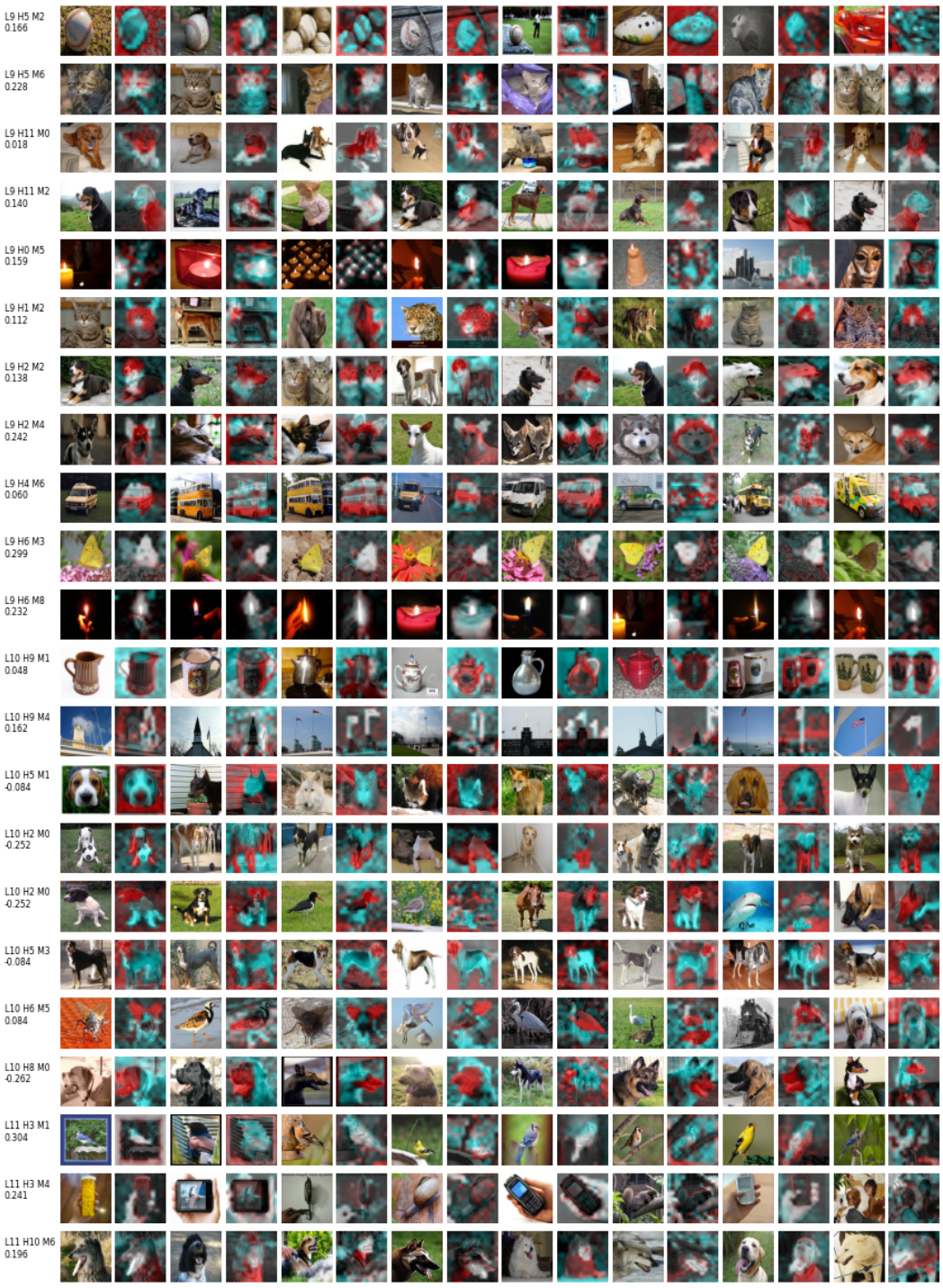

Figure S10: Examples of semantic singular modes in dino-vitb16 (part 3).

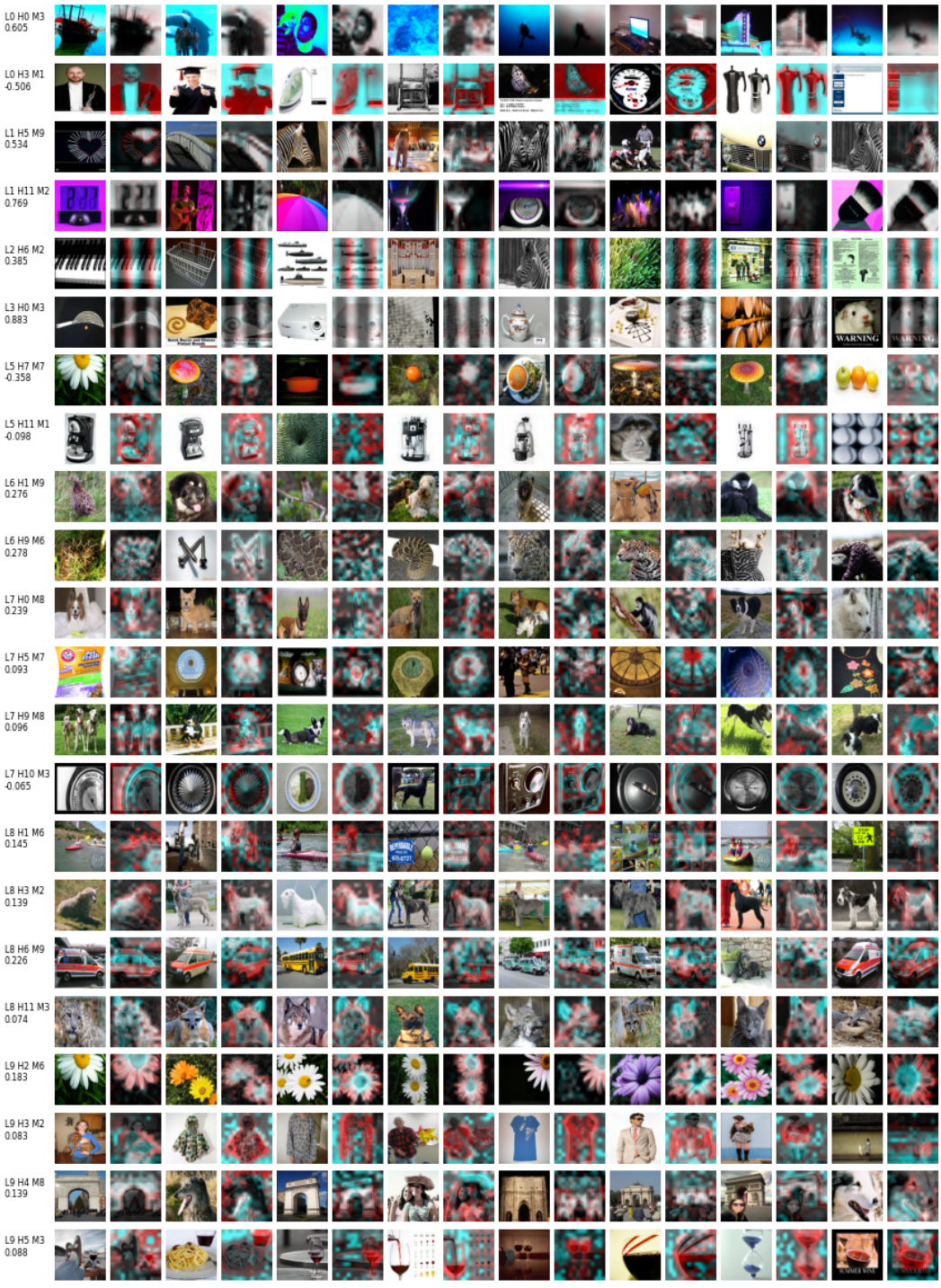

Figure S11: Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 1).

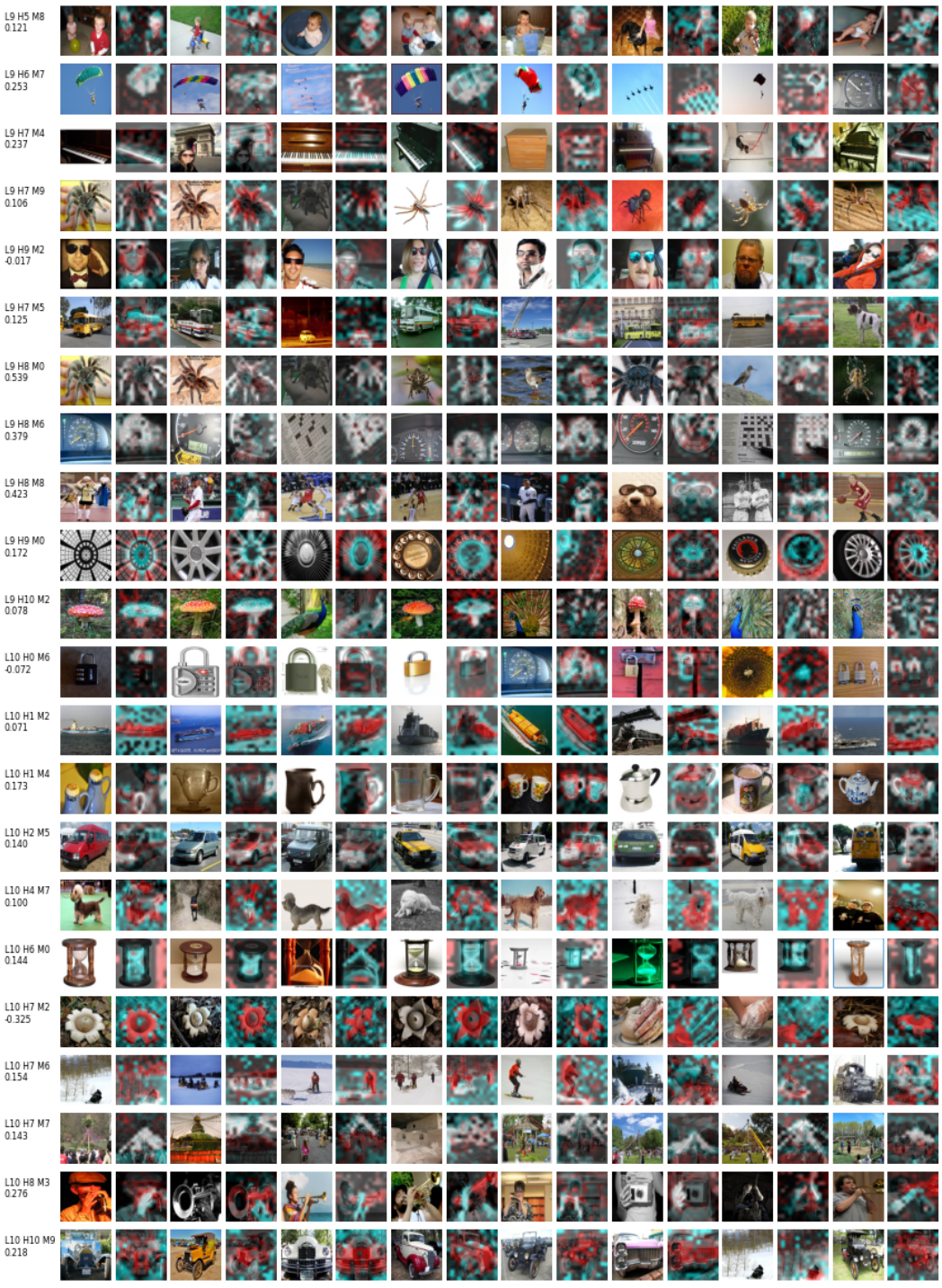

Figure S12: Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 2).

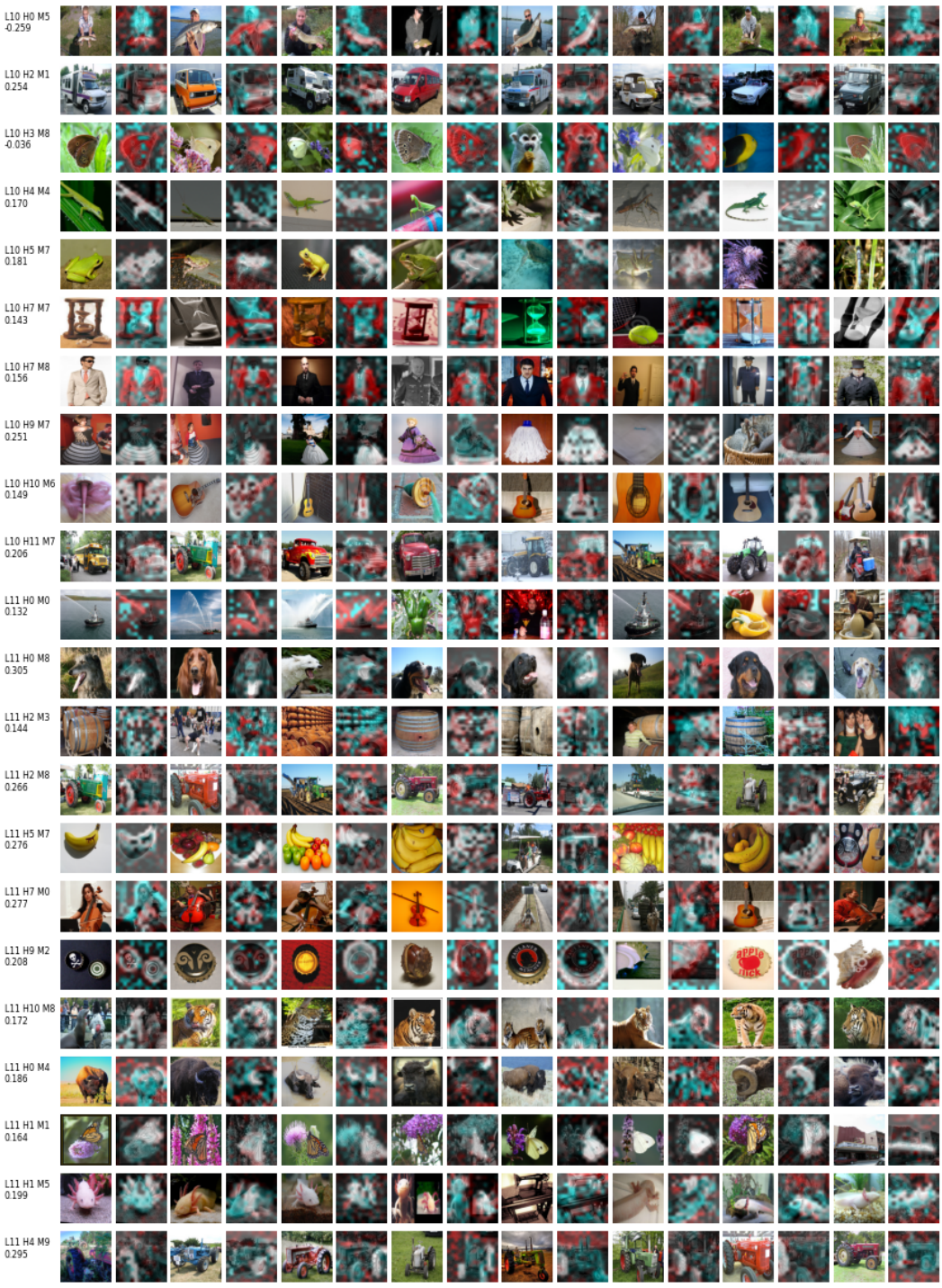

Figure S13: Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 3).

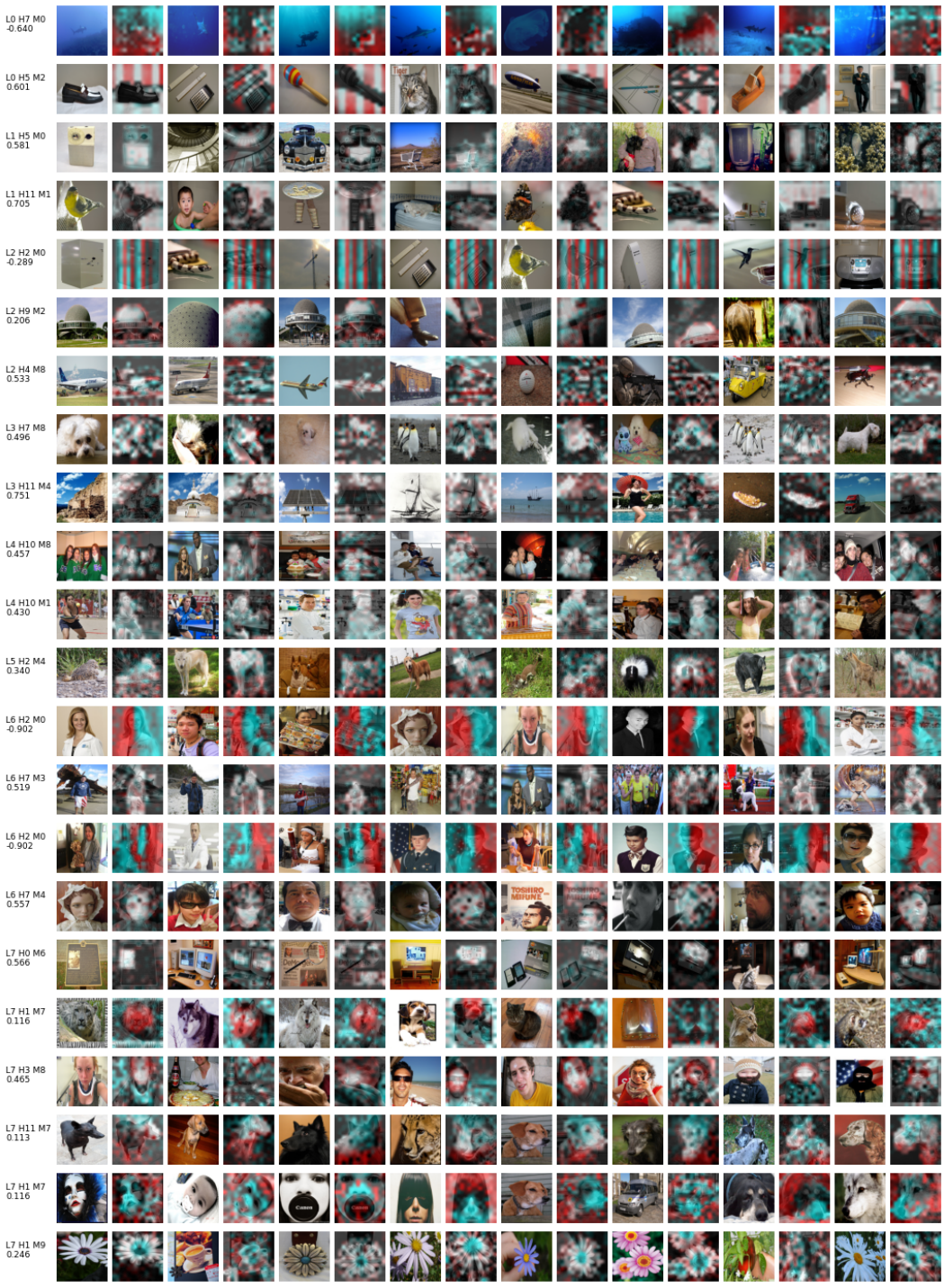

Figure S14: Examples of semantic singular modes in clip-vit-base-patch16 (part 1).

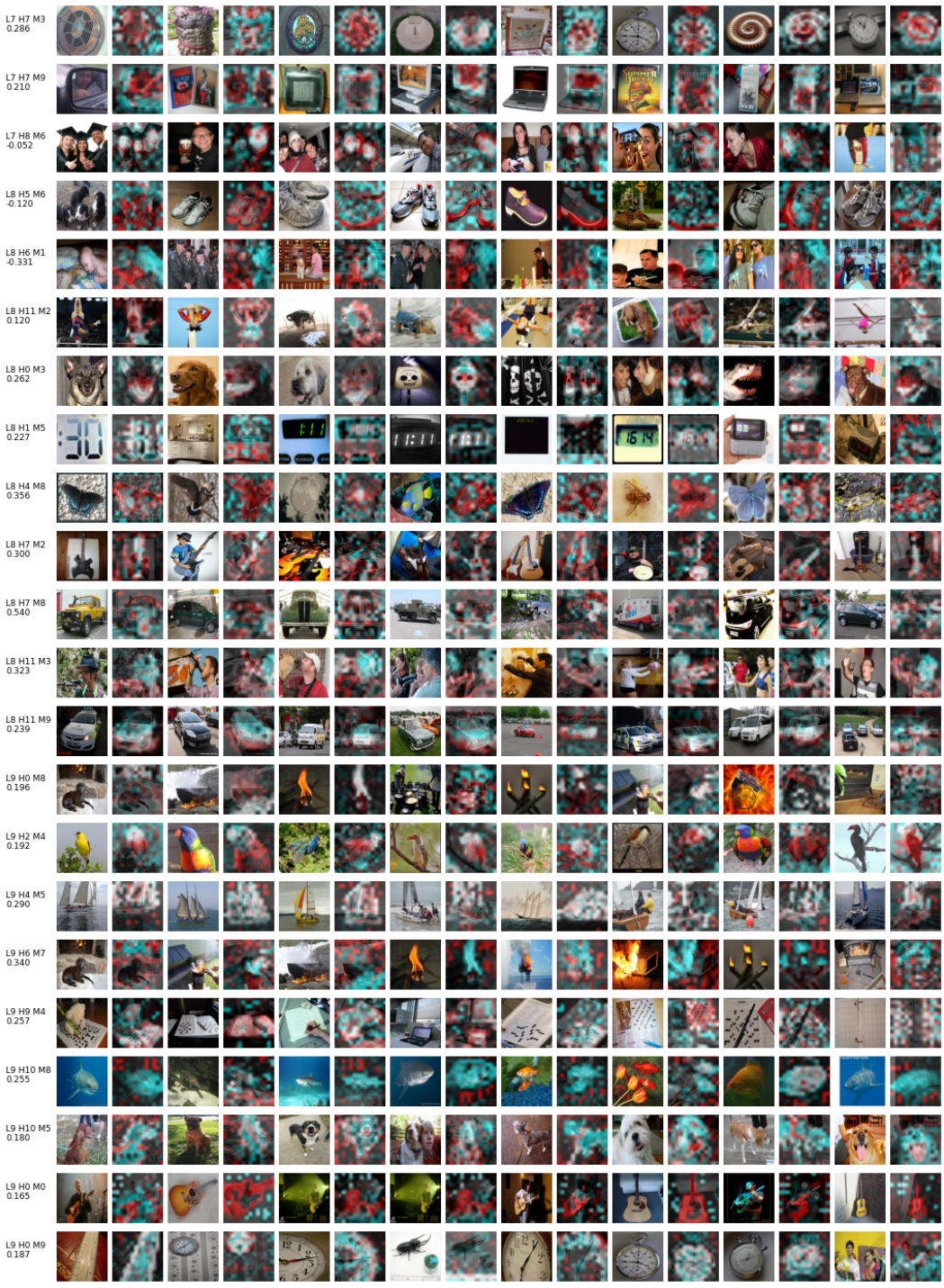

Figure S15: Examples of semantic singular modes in clip-vit-base-patch16 (part 2).

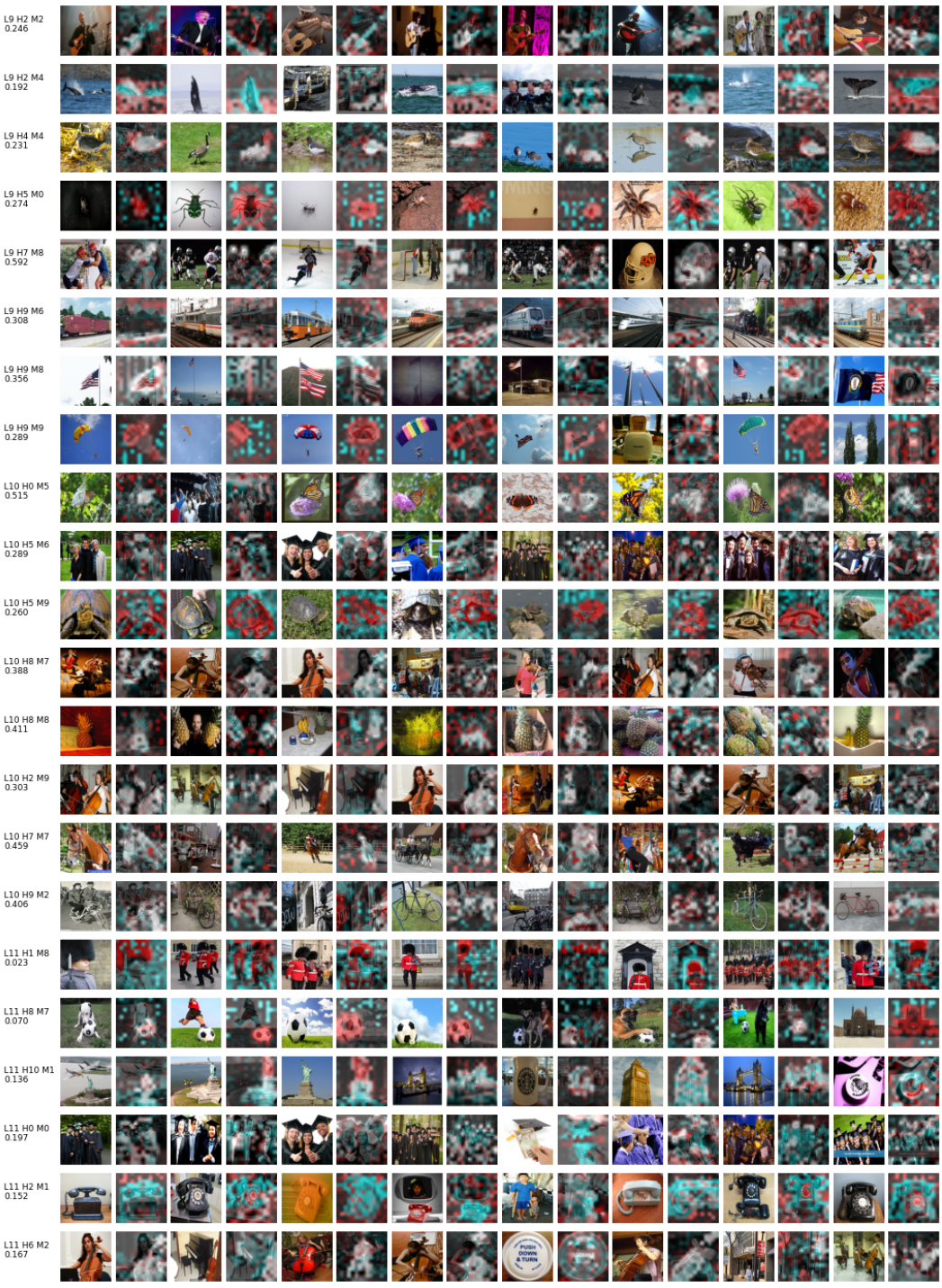

Figure S16: Examples of semantic singular modes in clip-vit-base-patch16 (part 3).

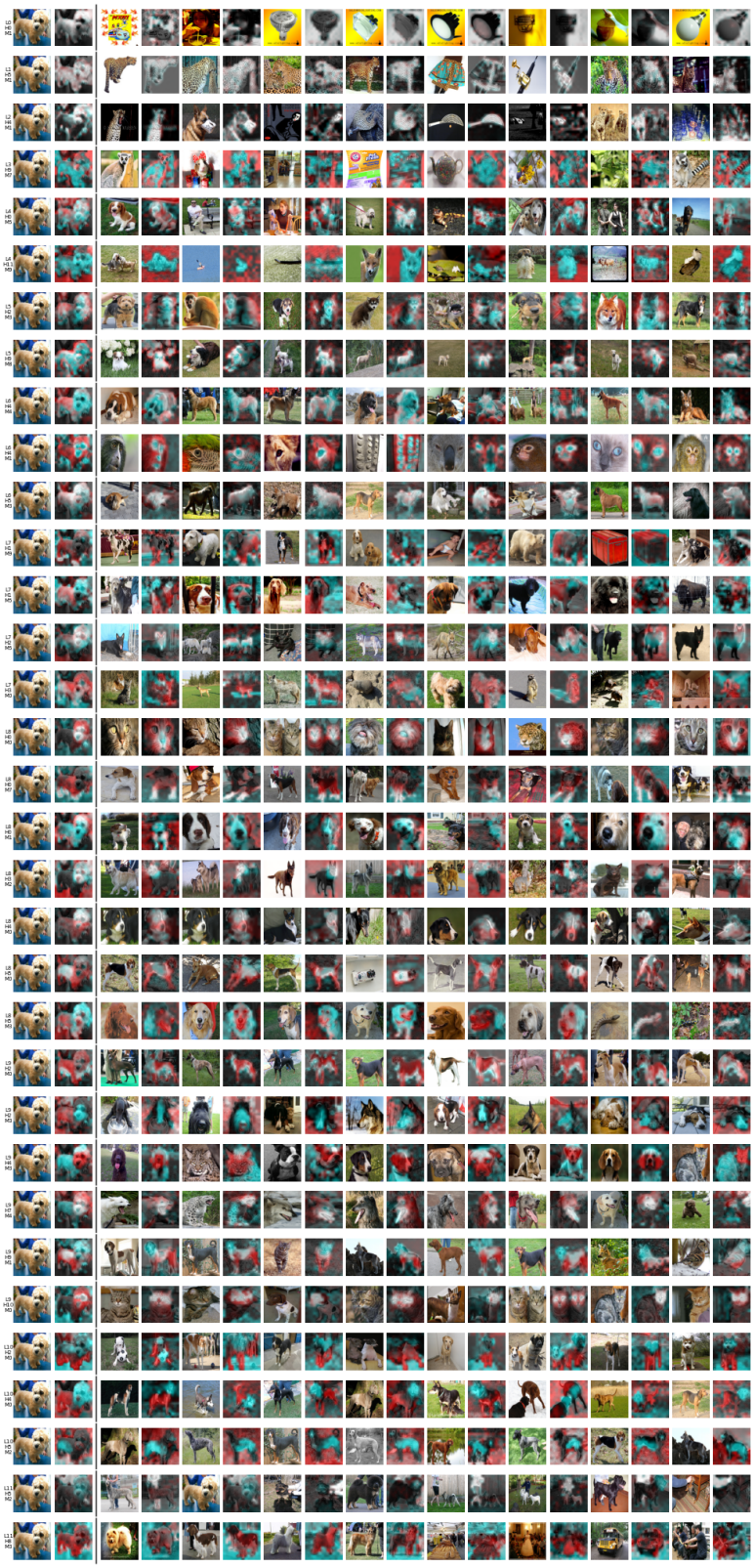

Figure S17: Singular mode maps of a dog image in dino-vitb16. We hand-pick modes to show the variety of information interactions within this image. The left two columns are the original image and corresponding singular mode maps. Other columns are the top 8 images that induce the highest attention through the corresponding mode.

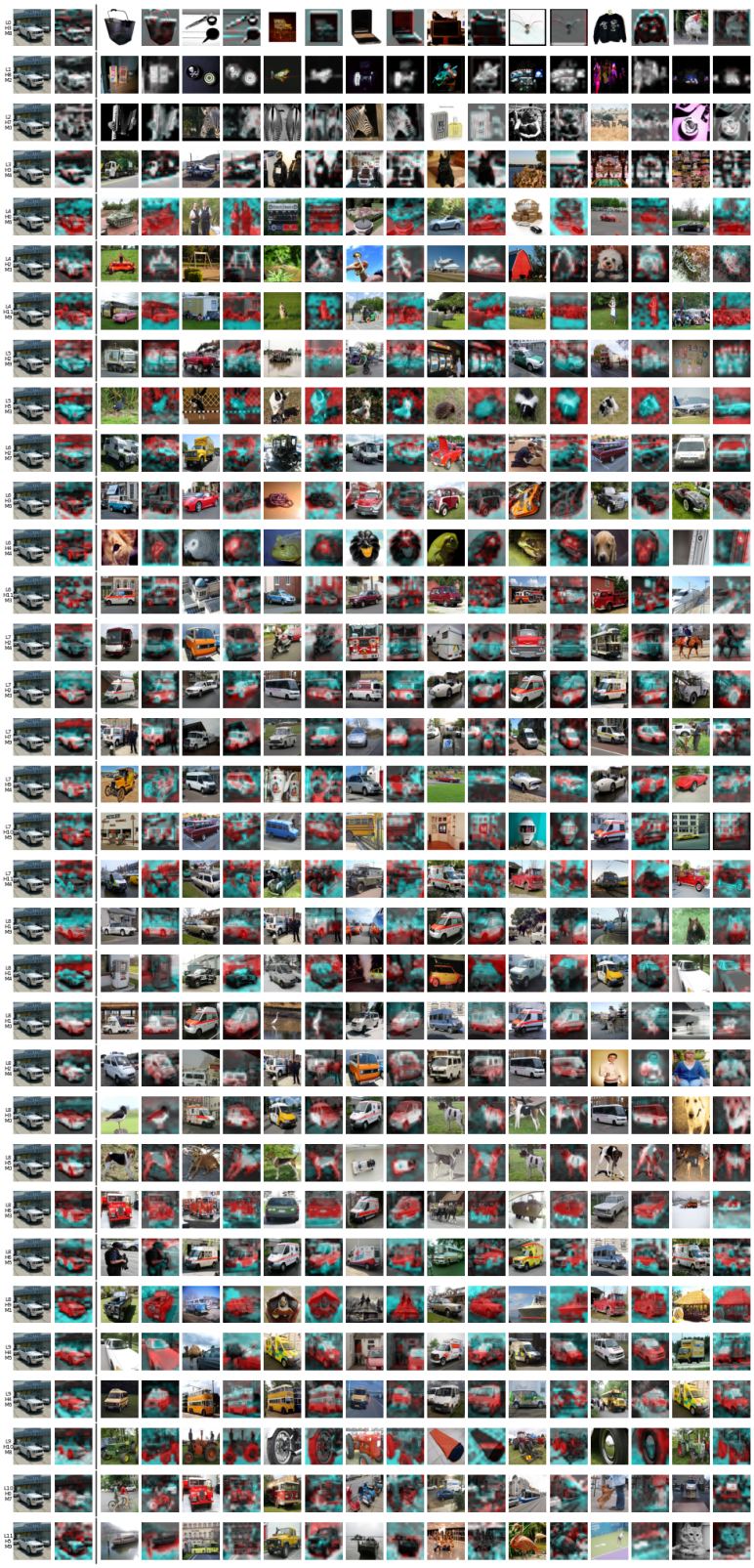

Figure S18: Singular mode maps of a car image in dino-vitb16. We hand-pick modes to show the variety of information interactions within this image. The left two columns are the original image and corresponding singular mode maps. Other columns are the top 8 images that induce the highest attention through the corresponding mode.

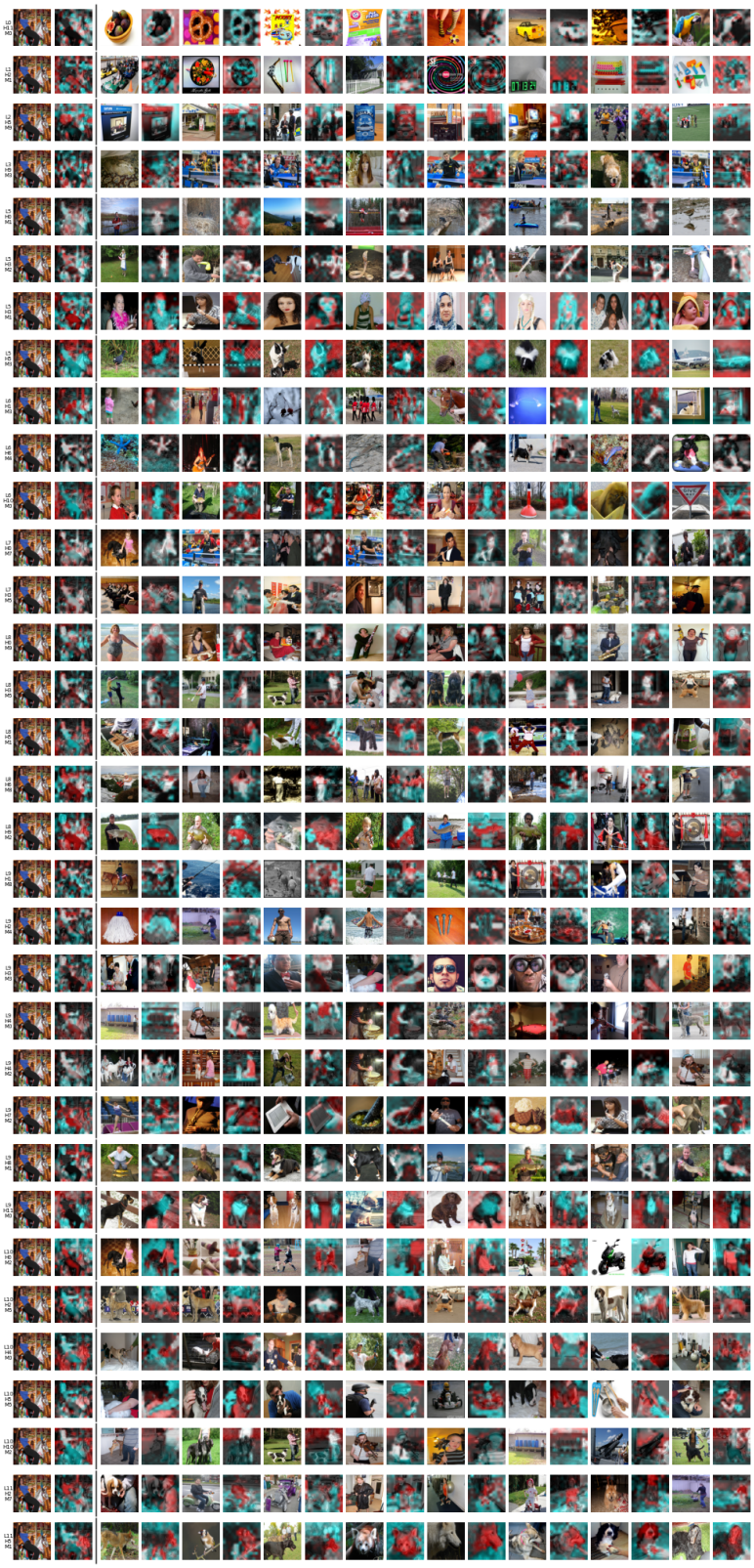

Figure S19: Singular mode maps of a human image in dino-vitb16. We hand-pick modes to show the variety of information interactions within this image. The left two columns are the original image and corresponding singular mode maps. Other columns are the top 8 images that induce the highest attention through the corresponding mode.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract accurately reflects the contributions and scope. We explicitly highlight the main contributions in the introduction.

   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: See section 6 (Limitation).

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?

   Answer: [No]

   Justification: We don't have theoretical results, but we write out in detail our use of the SVD to interpret query-key interactions.

   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: Our experiments use publicly available datasets, and the experiment details are provided in the paper.

   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**

   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: The code will be made publicly available upon acceptance. The online repository is being finalized for readability.

   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: All experimental details are provided. Our experiments do not require training a new model.

   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: [No]

   Justification: For readability, we do not draw errors in the main figures. We added distribution plots with confidence intervals in supplementary figures 1 and 2.

   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).
   - 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: [No]

   Justification: Our experiments do not require compute resources beyond a personal computer with a GPU.

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: All image datasets used in the experiments are publicly available. We use the Imagenet version available on Hugging Face.

   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: See the end of the discussion 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.
    - 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: We use existing models and datasets. We do not see high risk in our analysis.

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 cited all the datasets and models used in the paper. We noted that the Imagenet and models were obtained from Hugging Face.

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.
- 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: The 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: The 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: The paper does not involve research with human subjects and therefore does not include an IRB.

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.
- 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.