Dissecting Query-Key Interaction in Vision Transformers

Xu Pan [1] **Aaron Philip** [2] **Ziqian Xie** [3] **Odelia Schwartz** [1]

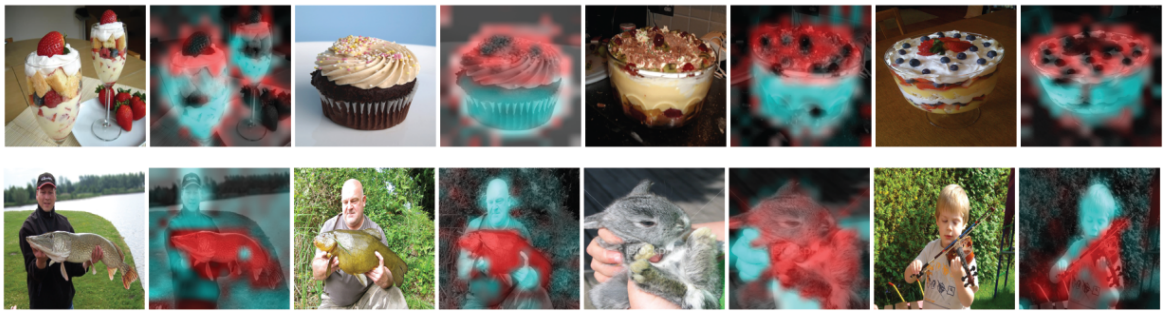

Figure 1. We propose a new way to study query-key interactions via the Singular Value Decomposition. Many of the modes (i.e. pairs of singular vectors corresponding to the query and the key respectively), are semantic. Two example modes are shown. Top row: ViT layer 8 head 7 mode 2. Bottom row: DINO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

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 use the Singular Value Decomposition to dissect the query-key interaction (i.e. $\mathbf{W}_q{}^\top \mathbf{W}_k$). We find that 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.

1. Introduction

Vision transformers (ViTs) are a family of models that have significantly advanced the computer vision field in recent years (Dosovitskiy et al., 2021). The core computation of ViTs, self-attention, is designed to promote interactions between tokens corresponding to relevant image features (Dosovitskiy et al., 2021). But this mechanism 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 (Dosovitskiy et al., 2021; Caron et al., 2021; Chen et al., 2022). However, observations in the language domain suggest that self-attention contextualizes tokens, such that the same token has different meanings in different contexts(Ethayarajh, 2019a). 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 (Cavanaugh et al., 2002; Jones et al., 2002; Ziemba et al., 2018; Li, 1999; Itti & Koch, 2001; Clifford & Rhodes, 2005; Angelucci et al., 2017; Choung et al., 2021). Therefore, two

[1]Department of Computer Science, University of Miami, Coral Gables, FL [2]Michigan State University, East Lansing, MI [3]University of Texas Health Science Center at Houston, Houston, Texas. Correspondence to: Xu Pan <panxu001@gmail.com>, Odelia Schwartz <odelia@cs.miami.edu>.

Workshop on Mechanistic Interpretability at the 41st International Conference on Machine Learning, Vienna, Austria.

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 (Geva et al., 2021; Bills et al., 2023; Bricken et al., 2023; Dar et al., 2023; Radhakrishnan et al., 2024; Ghiasi et al., 2022). A general finding is that embeddings in feedforward layers (i.e. MLP layers) are more semantically interpretable than in self-attention layers (Geva et al., 2021; Ghiasi et al., 2022). 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 (Bricken et al., 2023). 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. Specifically, 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.
- 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 progress towards 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.

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 (Dosovitskiy et al., 2021). 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 (Caron et al., 2021). 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 (Oquab et al., 2023; Wang et al., 2022). Other research analyzing 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 (Mehrani & Tsotsos, 2023). 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 like grouping and contextualization.

Contextualization Our study is inspired by contextual effects in visual neuroscience, in which neural responses are modulated by the surrounding context (Angelucci et al., 2017; Cavanaugh et al., 2002; Ziemba et al., 2018). 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 (Li, 1999; Coen-Cagli et al., 2015). Some of these biological surround contextual effects have been observed in convolutional neural networks (Marques et al., 2021; Pan et al., 2023). 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 (Ethayarajh, 2019a). 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 (Geva et al., 2021), research has primarily focused on finding feature axes in the feedforward layers, and to a lesser extent, in self-attention layers. Bills et al. proposed a gradient-based optimization method to find explainable directions in LLMs (Bills et al., 2023). Later, Bricken et al. proposed a simpler method of sparse dictionary learning (Bricken et al., 2023); though see also (Huben). These methods have not been applied to studies of ViTs. However, similar to the findings in LLMs, a ViT study found that feedforward layers have less mixed concepts and can generate interpretable feature visualizations (Ghiasi et al., 2022).

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 (Siméoni et al., 2021; Amir et al., 2021; Adeli et al., 2023). A study proposed that the singular value decomposition of the weight matrix is a natural way to find feature directions in any neural network (Radhakrishnan et al., 2024). 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 (Millidge & Black, 2022). 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 (Lieberum et al., 2023).

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 (Dar et al., 2023). However, this approach does not find features beyond the standard basis of the query and key embeddings. 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 (Kotseruba et al., 2019), namely the Odd-One-Out (O3) dataset (Mehrani & Tsotsos, 2023). This dataset was also used by Mehrami et al (Mehrani & Tsotsos, 2023) 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 token of one category (target or distractors) covers more of the same category, different category, or background.

We chose to study 12 different ViT models from 4 families: the original ViT (Dosovitskiy et al., 2021), DeiT which uses distillation to learn from a teacher model (Touvron et al., 2021), DINO which is trained in a self-supervised way (Caron et al., 2021), and CLIP which is jointly trained with a text encoder (Radford et al., 2021).

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. However, the attention on the same objects is only dominant in early layers. In the deeper layers, there is a trend that attention gradually increases on the contextual features such as the background or different objects. Attention on the backgrounds surpasses that of attention on the same objects in the last few layers.

This result provides new evidence that self-attention considers contextual features as much or more than similar features in deeper layers. The self-attention only prefers the same

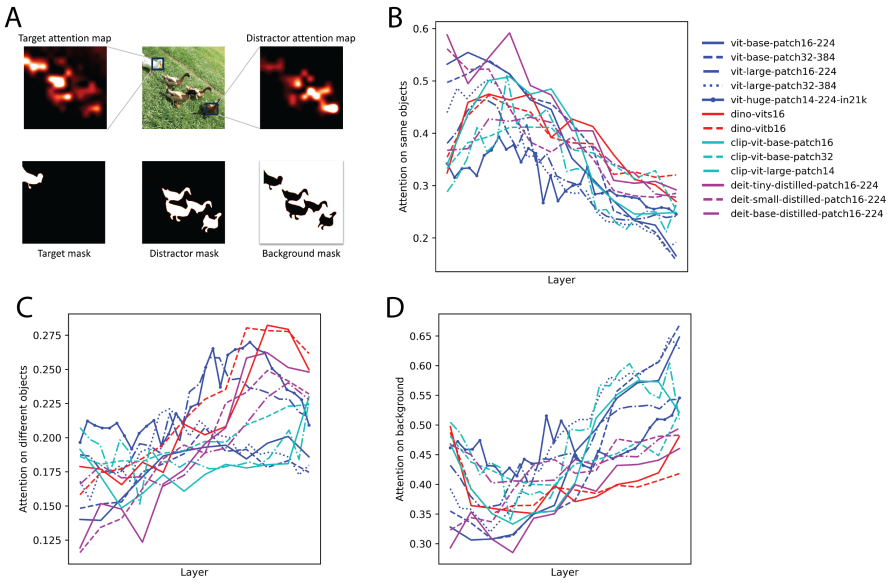

Figure 2. Attention preference in the Odd-One-Out (O3) dataset (Kotseruba et al., 2019). A. An example from the O3 dataset. Two tokens are chosen to correspond to the target and distractor in the image. Attention maps using two tokens as queries are computed. We examine the overlap between the attention 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). 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)

objects in early layers; in deeper layers, self-attention shifts to prefer contextual information. As far as the authors are aware, this finding has not been reported in previous ViT studies (Dosovitskiy et al., 2021; Caron et al., 2021; Wang et al., 2022; Mehrani & Tsotsos, 2023).

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 (Dosovitskiy et al., 2021):

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

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 (Elhage et al., 2021), 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

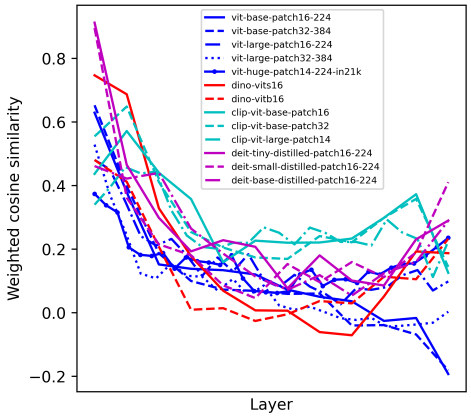

Figure 3. Cosine similarity between left and right singular vectors. The cosine similarity is computed per head and singular mode. The weighted average value of cosine similarity is computed with weights of corresponding singular values.

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.

It is known that embeddings in transformer models are to some extent anisotropic (Ethayarajh, 2019b; Liang et al., 2022; Godey et al., 2023), 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 S3) (though see other metrics (Rudman et al., 2021)). 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 S3).

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

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

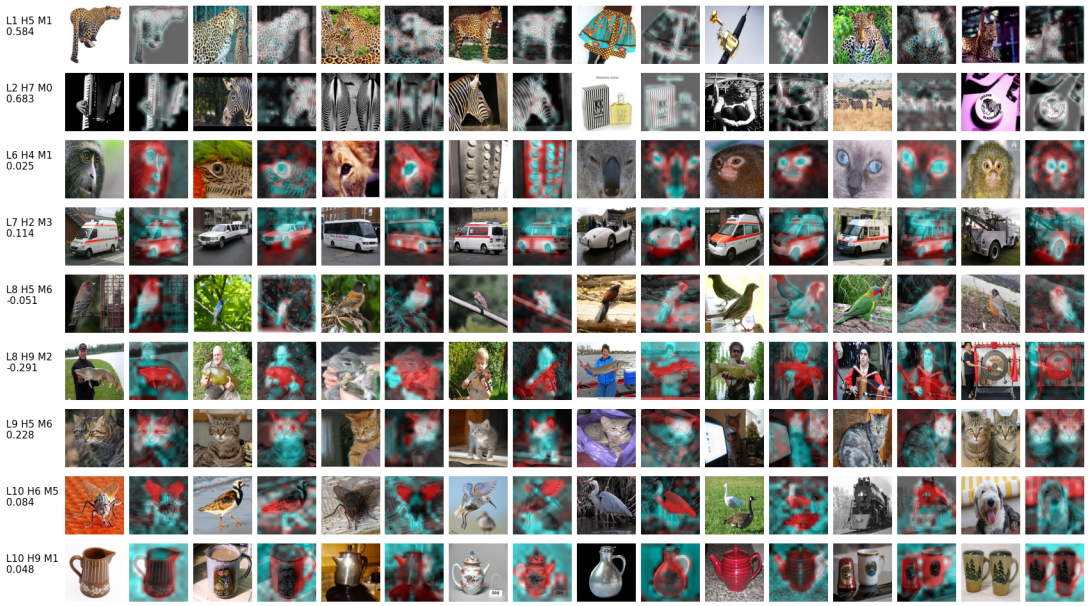

Figure 4. Examples of optimal attention images of singular modes and query and key map in dino-vitb16. Optimal attention images are found from the Imagenet validation set that induce the largest attention score (sorted by the product of the maximum of query map and maximum of key map). The red and cyan (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.

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 (Dosovitskiy et al., 2021; Caron et al., 2021; Oquab et al., 2023).

Some example modes (selected from the top 10 modes) from dino-vitb16 are shown in Fig. 4. For each mode, we choose the top 8 images in the Imagenet (Hugging Face version) (Russakovsky et al., 2015) 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 S4 - S15.

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.

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.

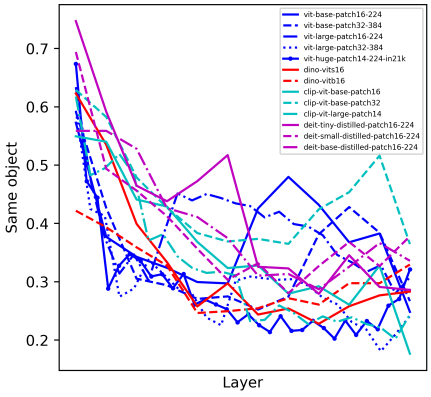

Figure 5. The probability that the left and right singular vectors highlight the same object in maximum attention images.

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 (Zhou et al., 2017). 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. 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.

References

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

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

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

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. 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.

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

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

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

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

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

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

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

Dar, G., Geva, M., Gupta, A., and Berant, J. Analyzing transformers in embedding space. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 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.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. 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.

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

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

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

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

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

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

Huben, R. 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].

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

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

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

Li, Z. 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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. Supplemental material

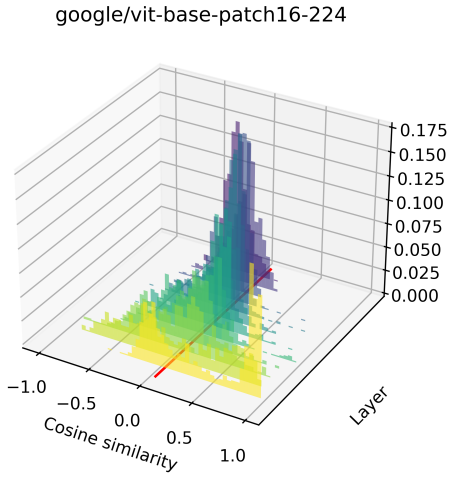

Figure S1. Histogram of cosine similarity between the left and right singular vector in ViT-base-patch16-224. The yellow layers are earlier layers; the blue layers are later layers. The red line indicates 95% confidence interval, which is calculated from embeddings sampled from a random distribution.

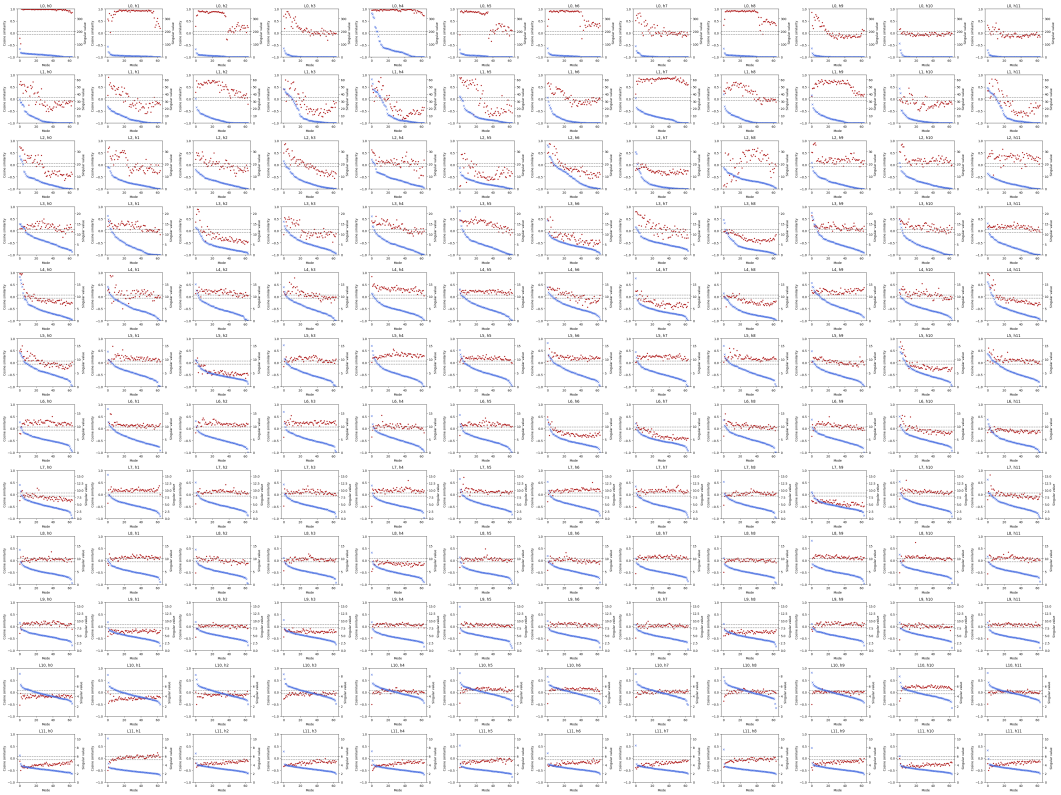

Figure S2. Singular value spectrum (blue) and cosine similarity (red) in ViT-base-patch16-224. Row number indicates layer number. Column number indicates head number. The dotted line indicates 95% confidence interval, which is calculated from embeddings sampled from a random distribution.

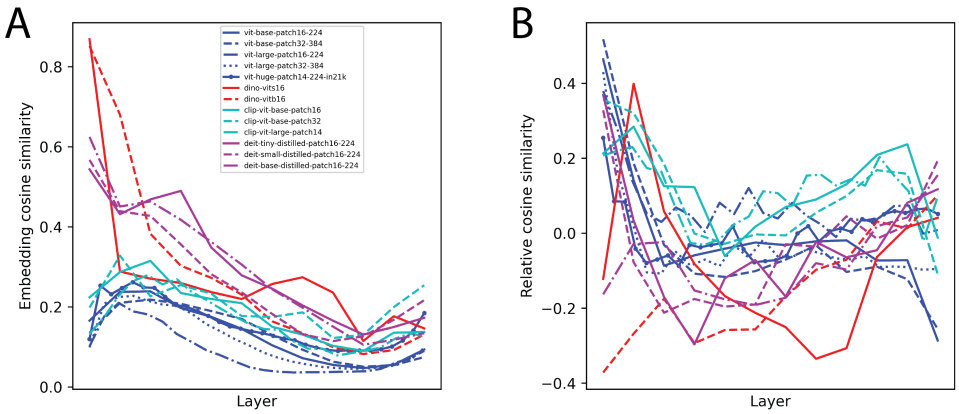

Figure S3. Anisotropy effects in ViTs. A. Averaged embedding cosine similarity between the center tokens of different images from the Imagenet validation set. Consisting with previous studies, the cosine similarities are all positive, which is referred to as anisotropy or cone effect. B. Considering 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.

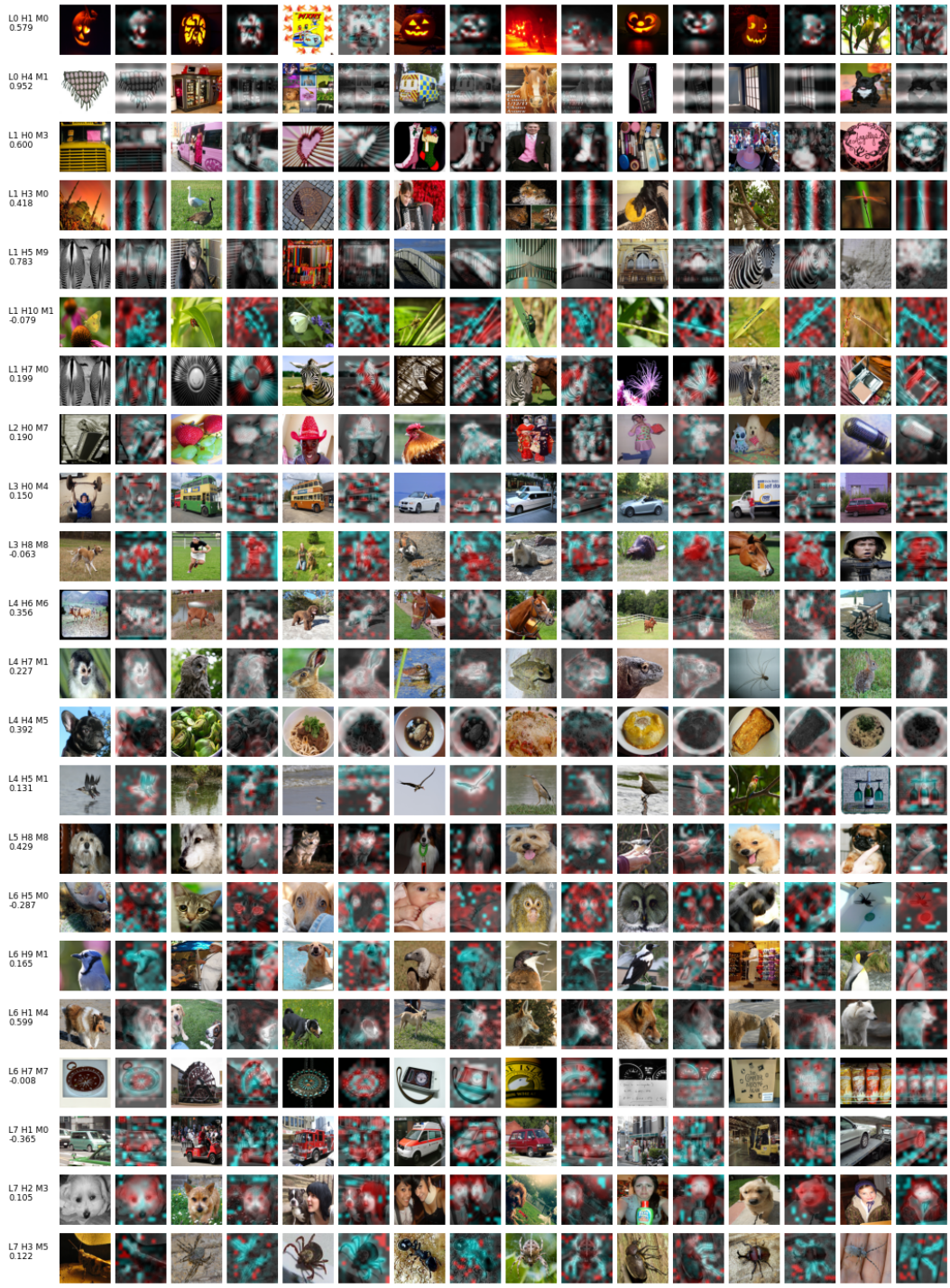

Figure S4. Examples of semantic singular modes in ViT-base-patch16-224 (part 1).

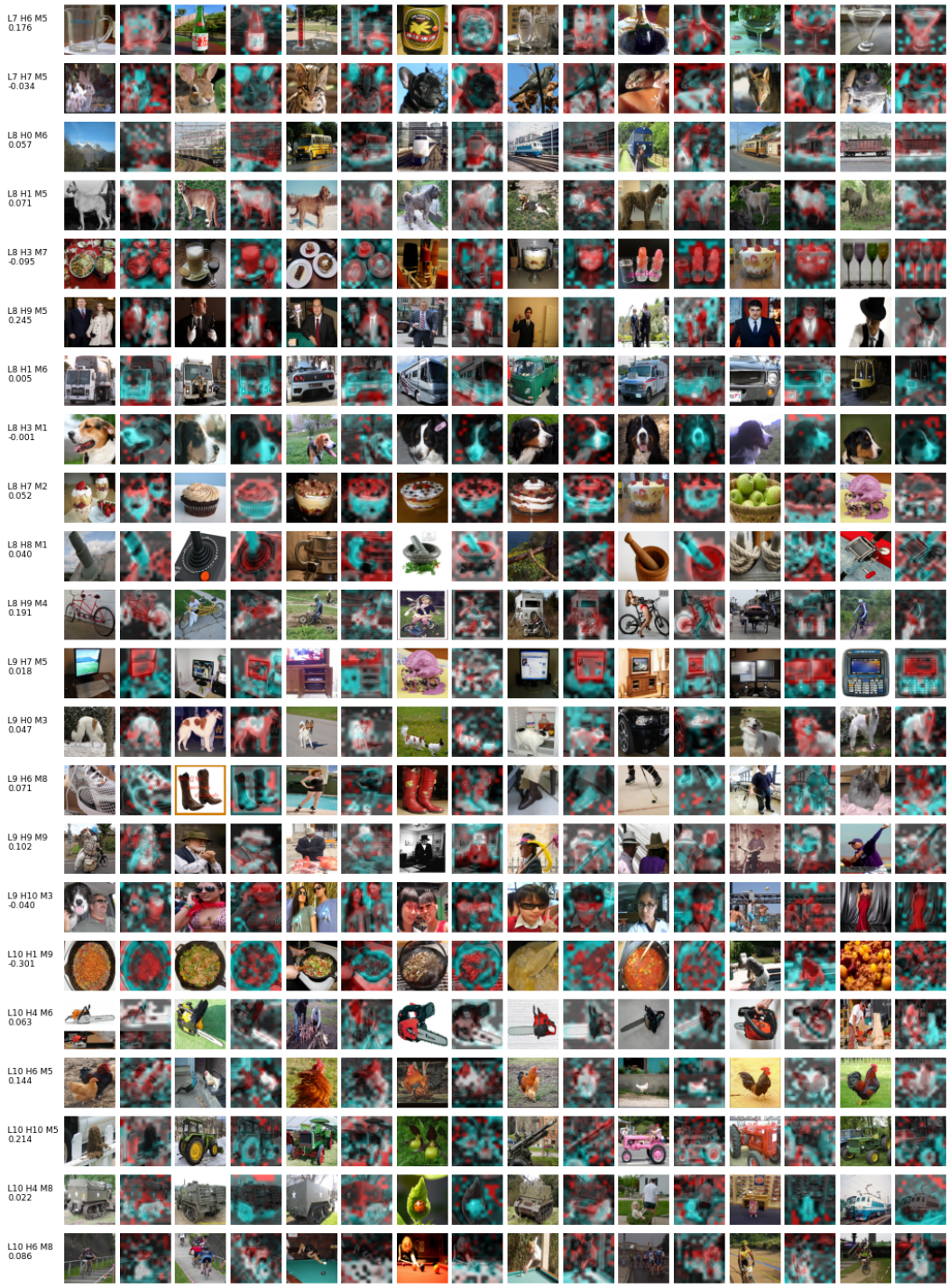

Figure S5. Examples of semantic singular modes in ViT-base-patch16-224 (part 2).

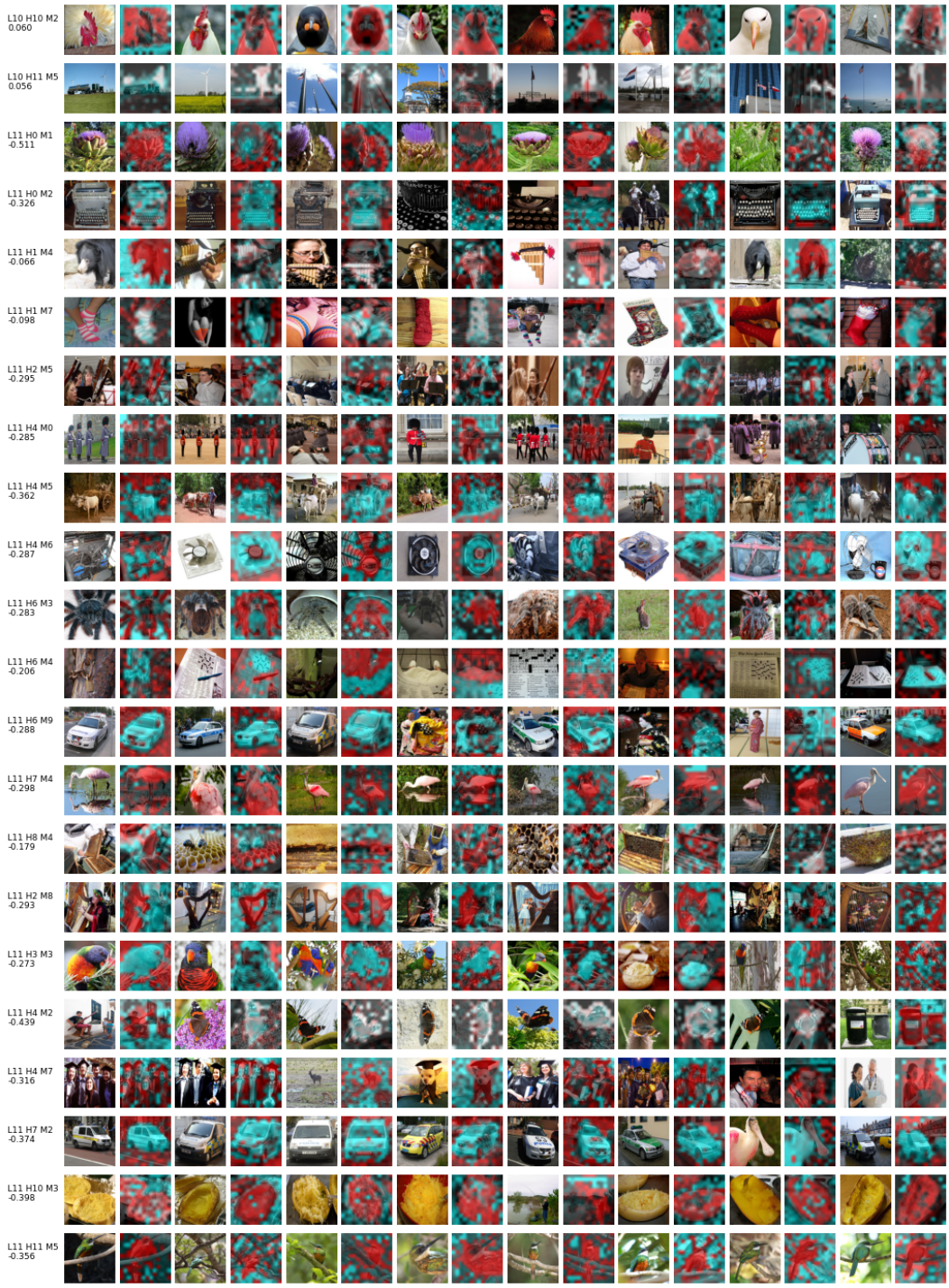

Figure S6. Examples of semantic singular modes in ViT-base-patch16-224 (part 3).

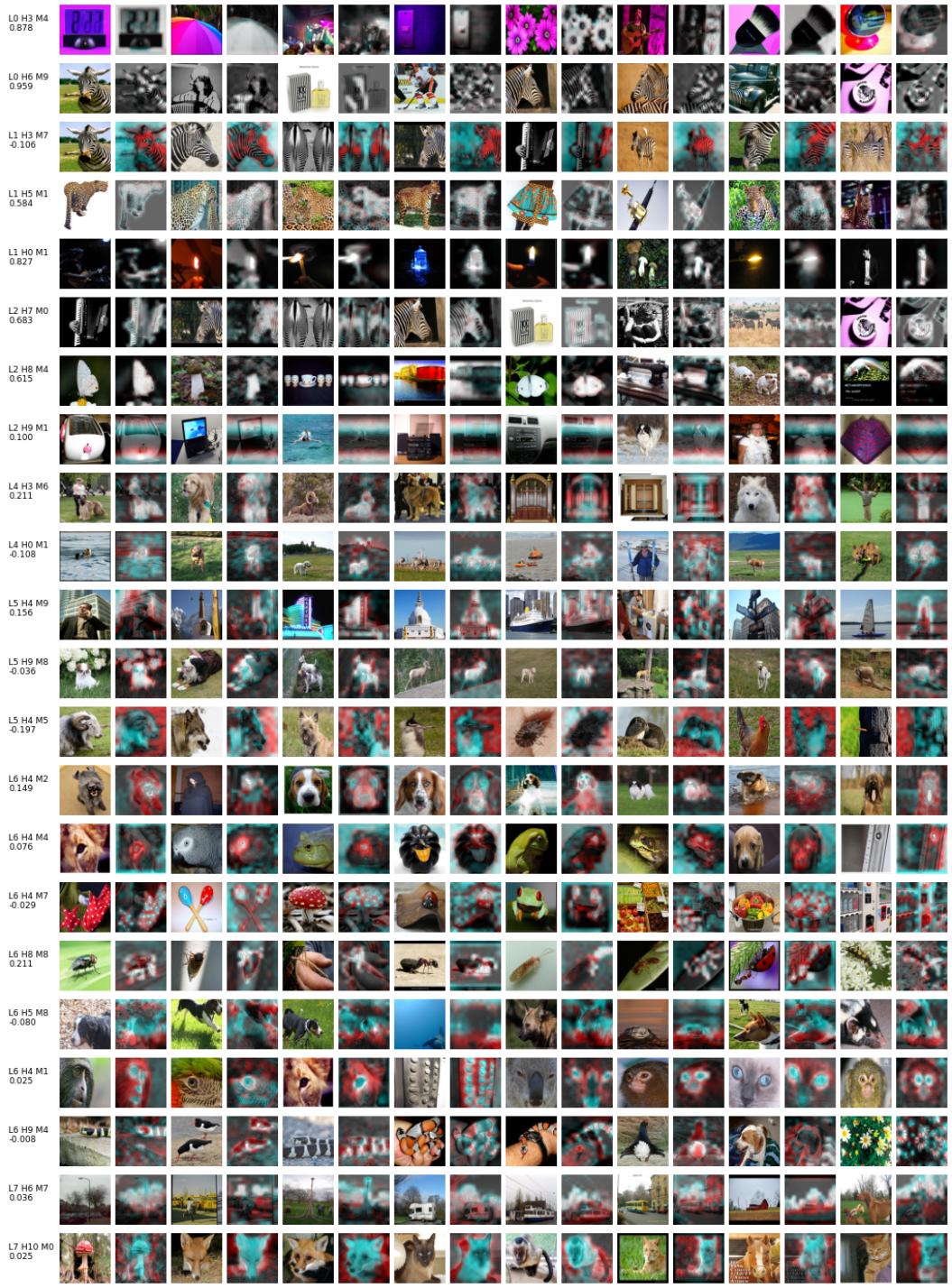

Figure S7. Examples of semantic singular modes in dino-vitb16 (part 1).

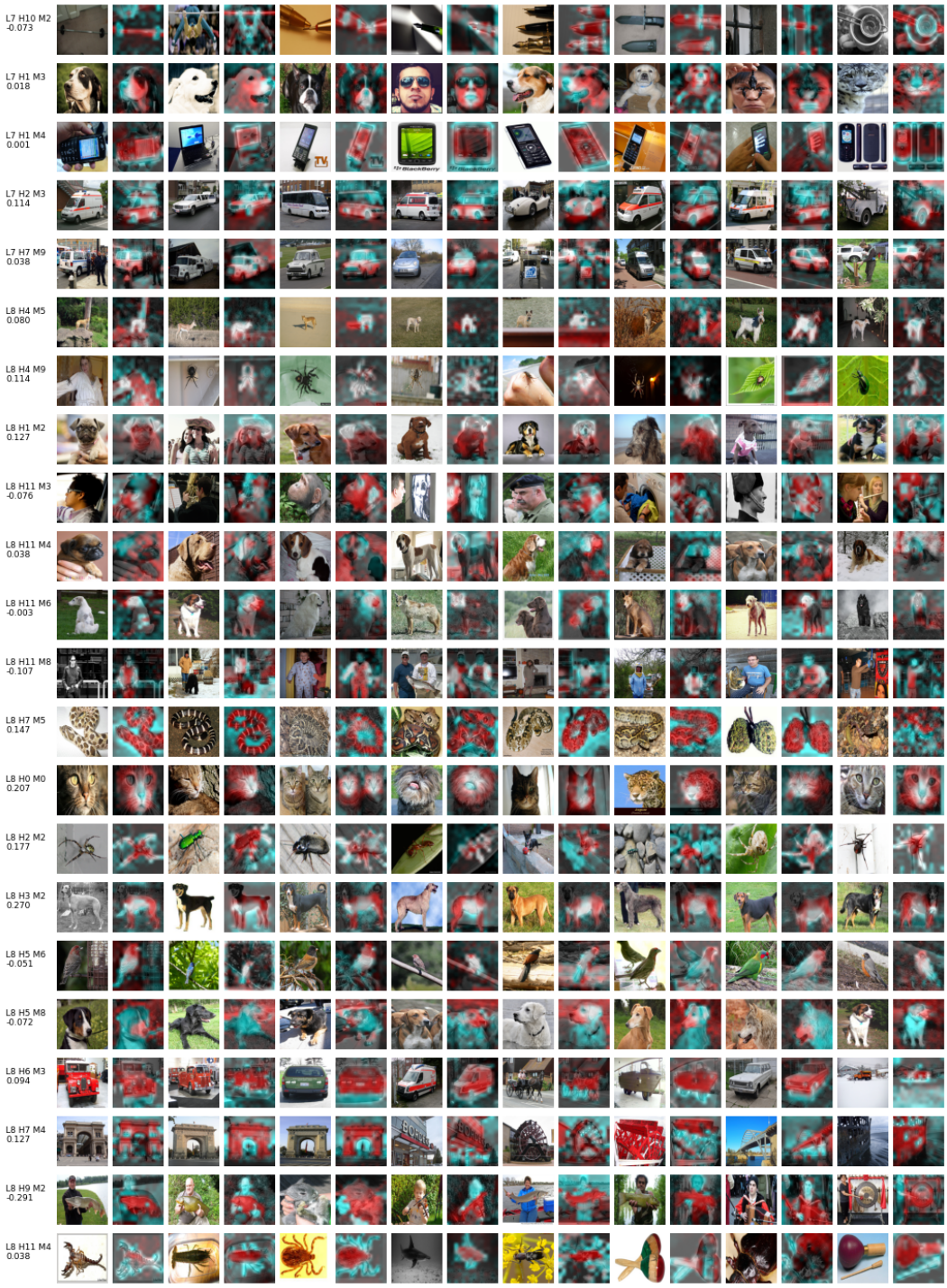

Figure S8. Examples of semantic singular modes in dino-vitb16 (part 2).

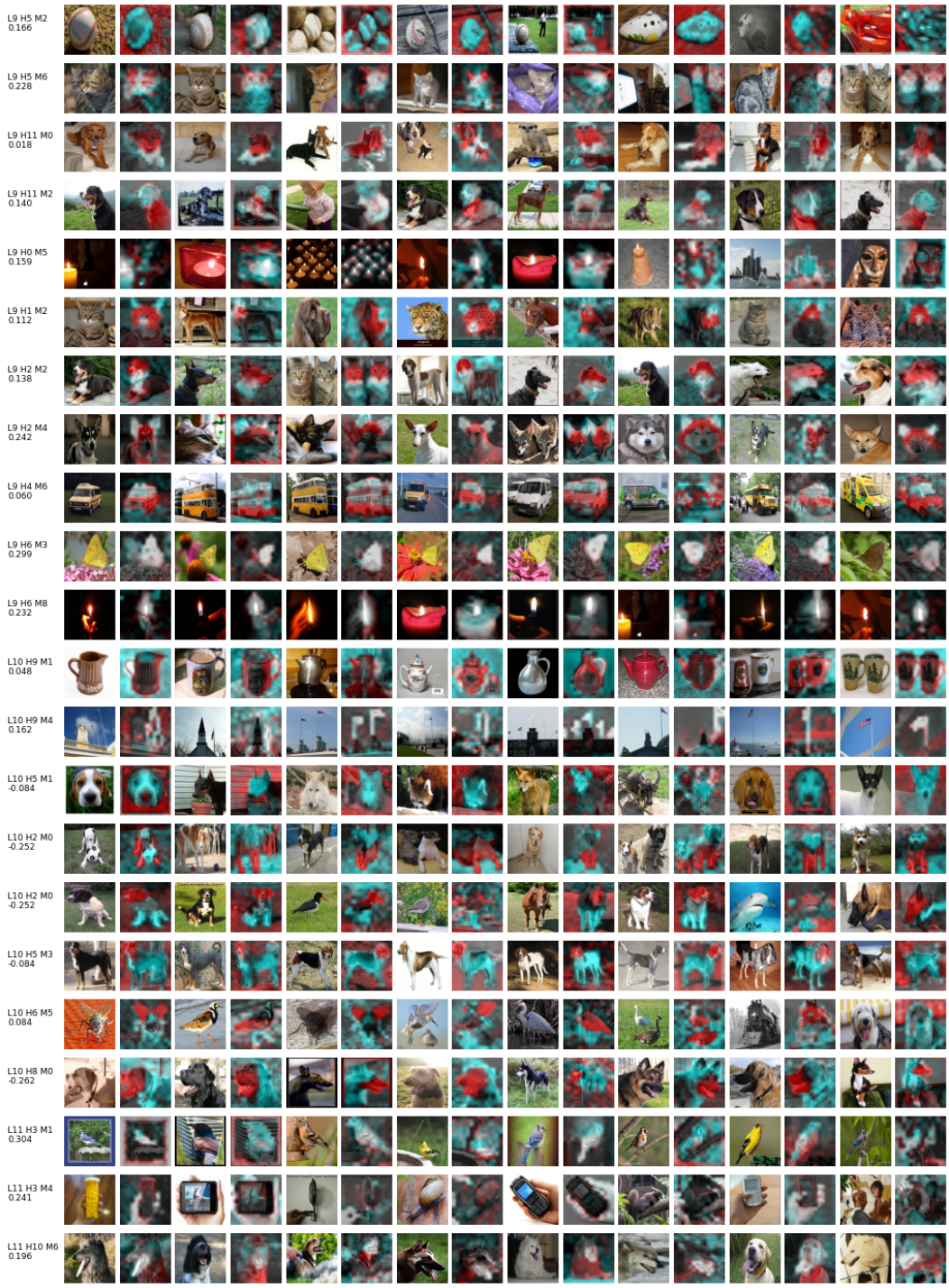

Figure S9. Examples of semantic singular modes in dino-vitb16 (part 3).

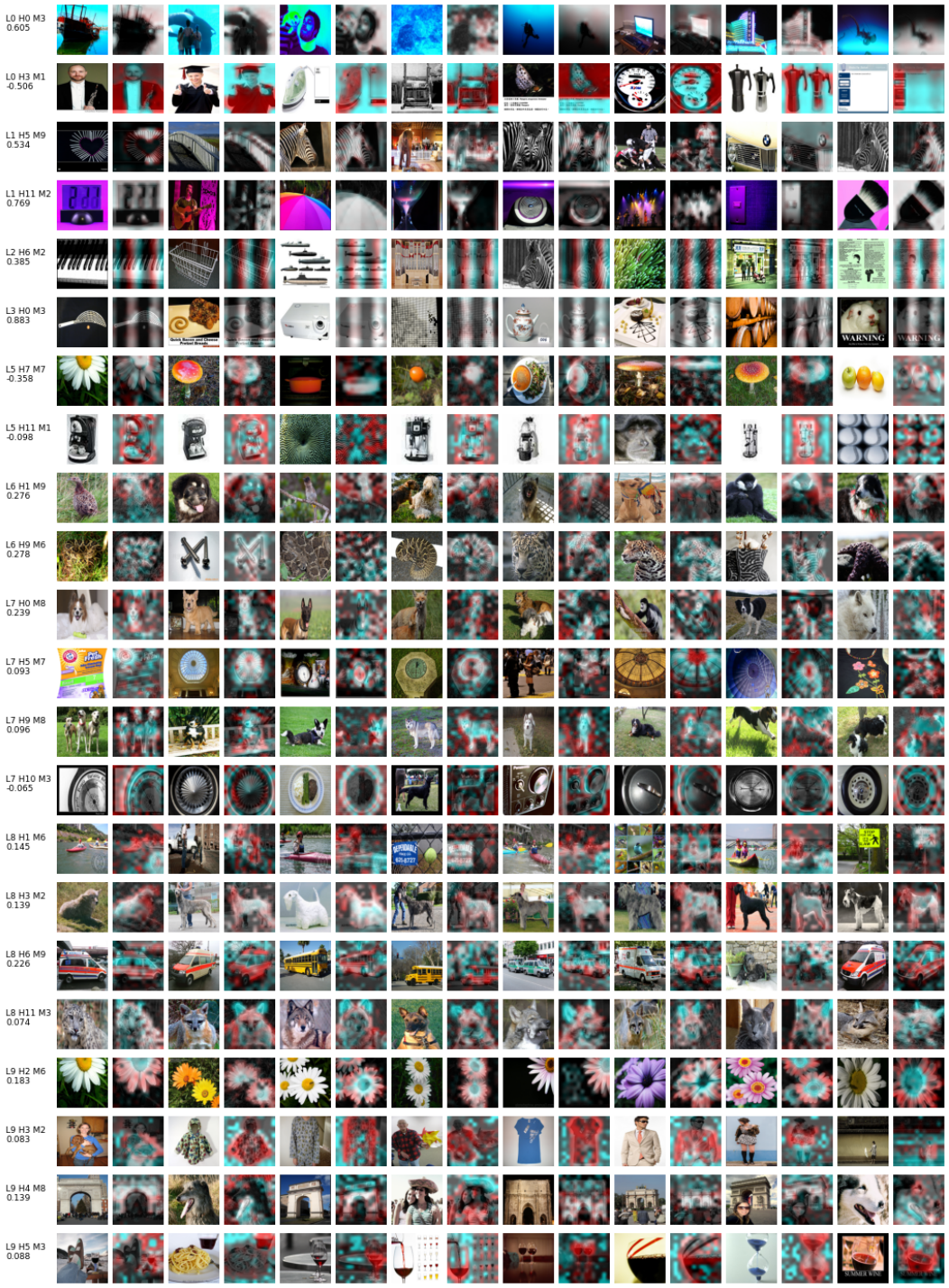

Figure S10. Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 1).

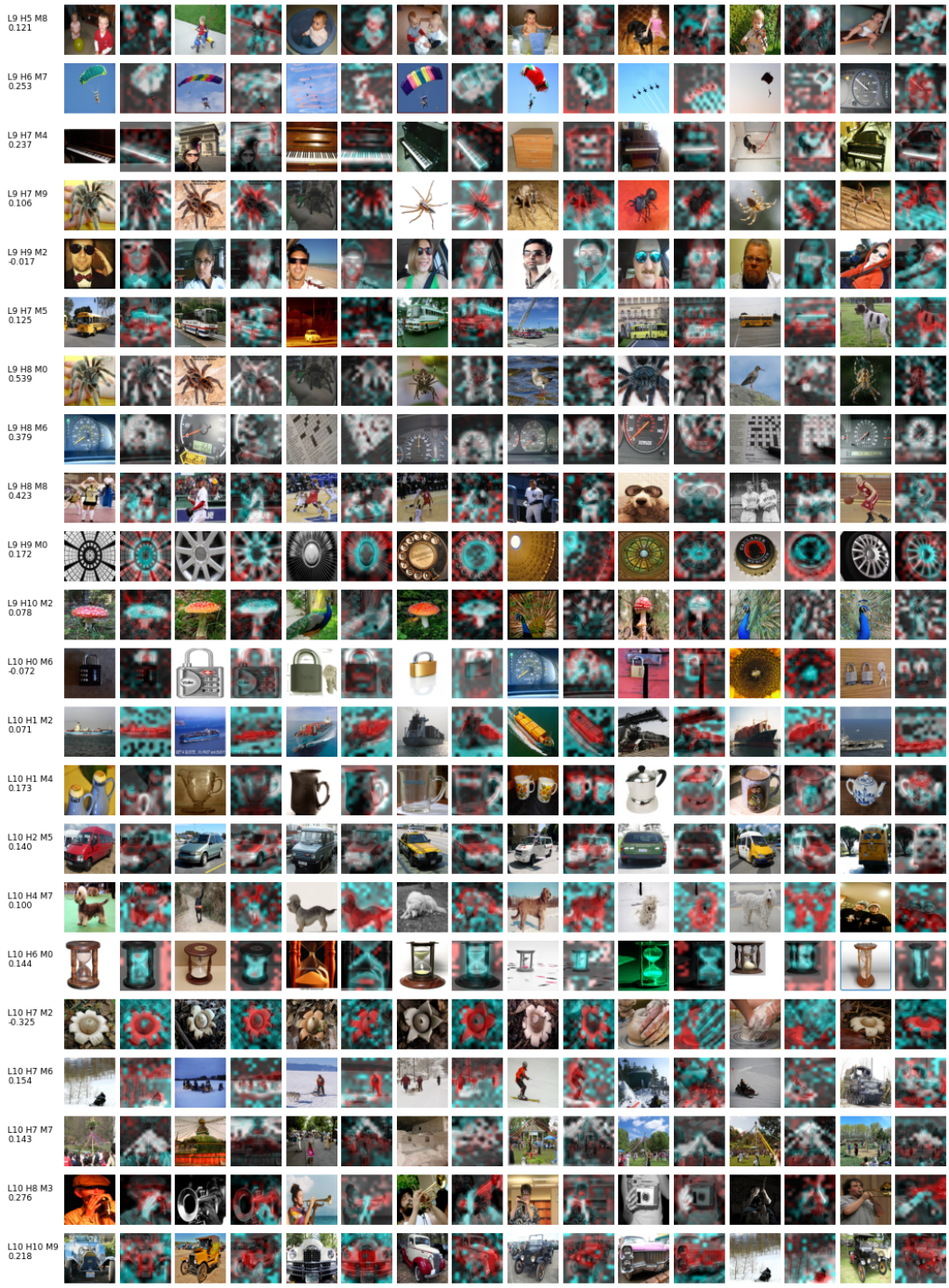

Figure S11. Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 2).

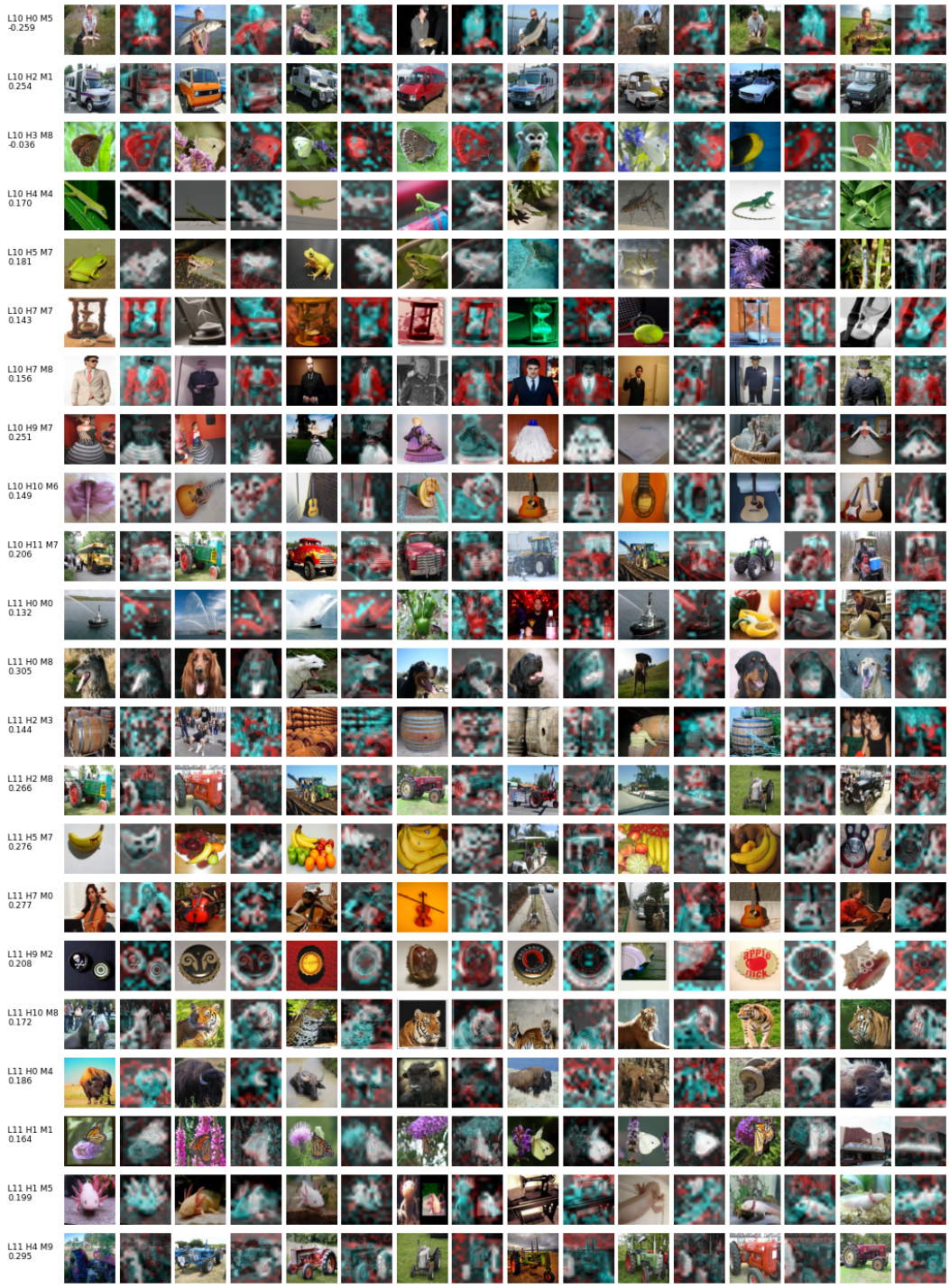

Figure S12. Examples of semantic singular modes in deit-base-distilled-patch16-224 (part 3).

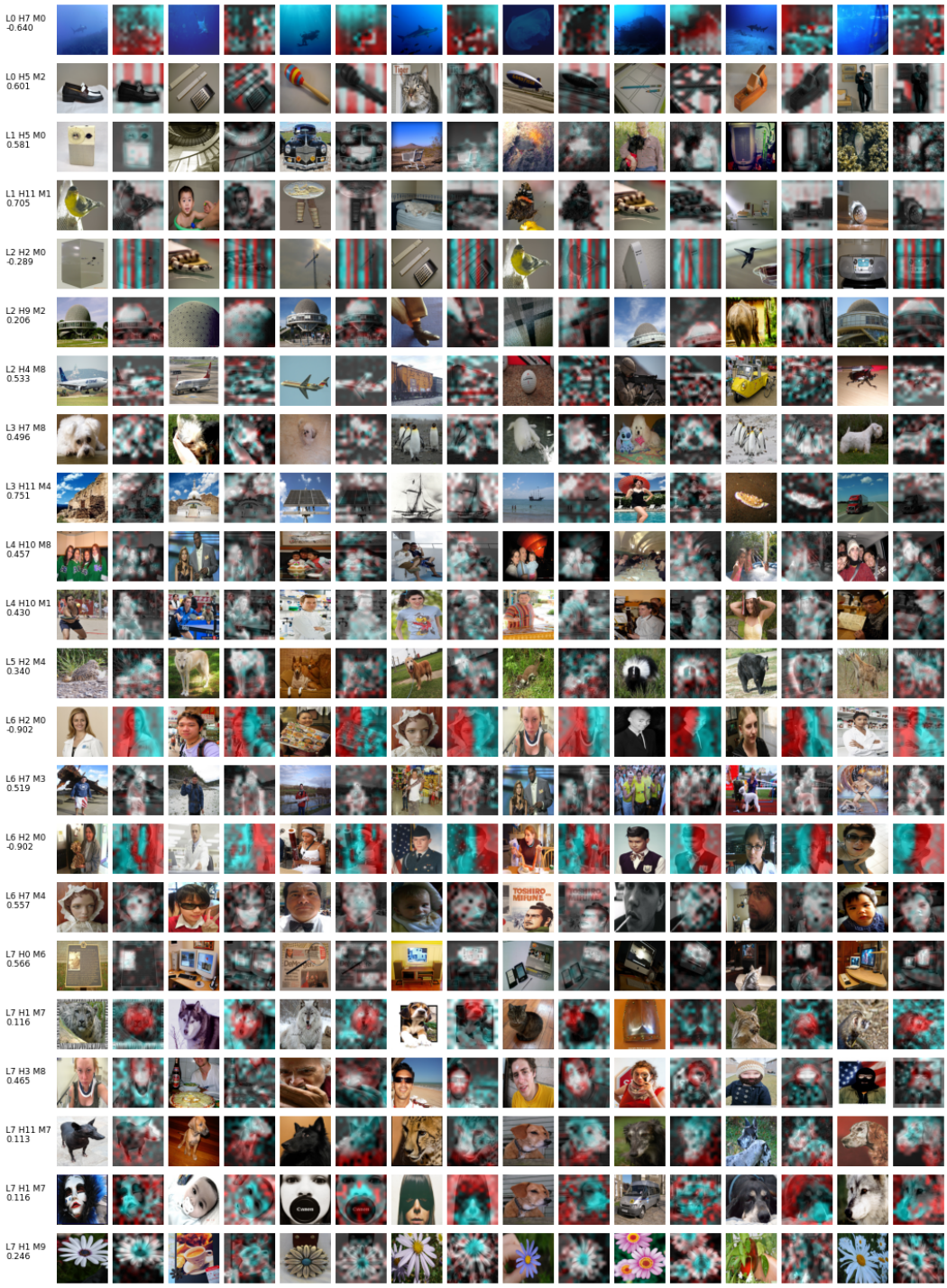

Figure S13. Examples of semantic singular modes in clip-vit-base-patch16 (part 1).

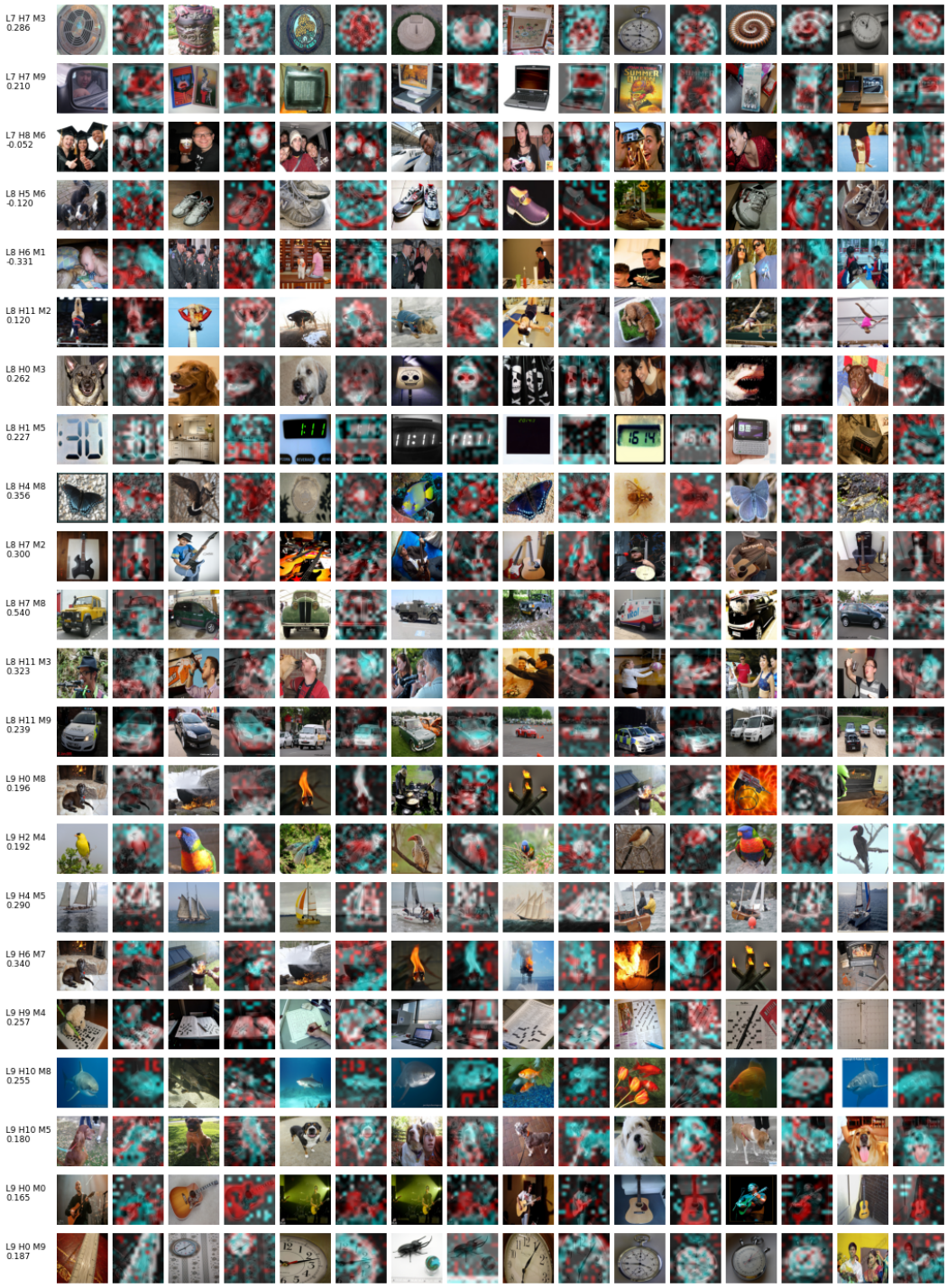

Figure S14. Examples of semantic singular modes in clip-vit-base-patch16 (part 2).

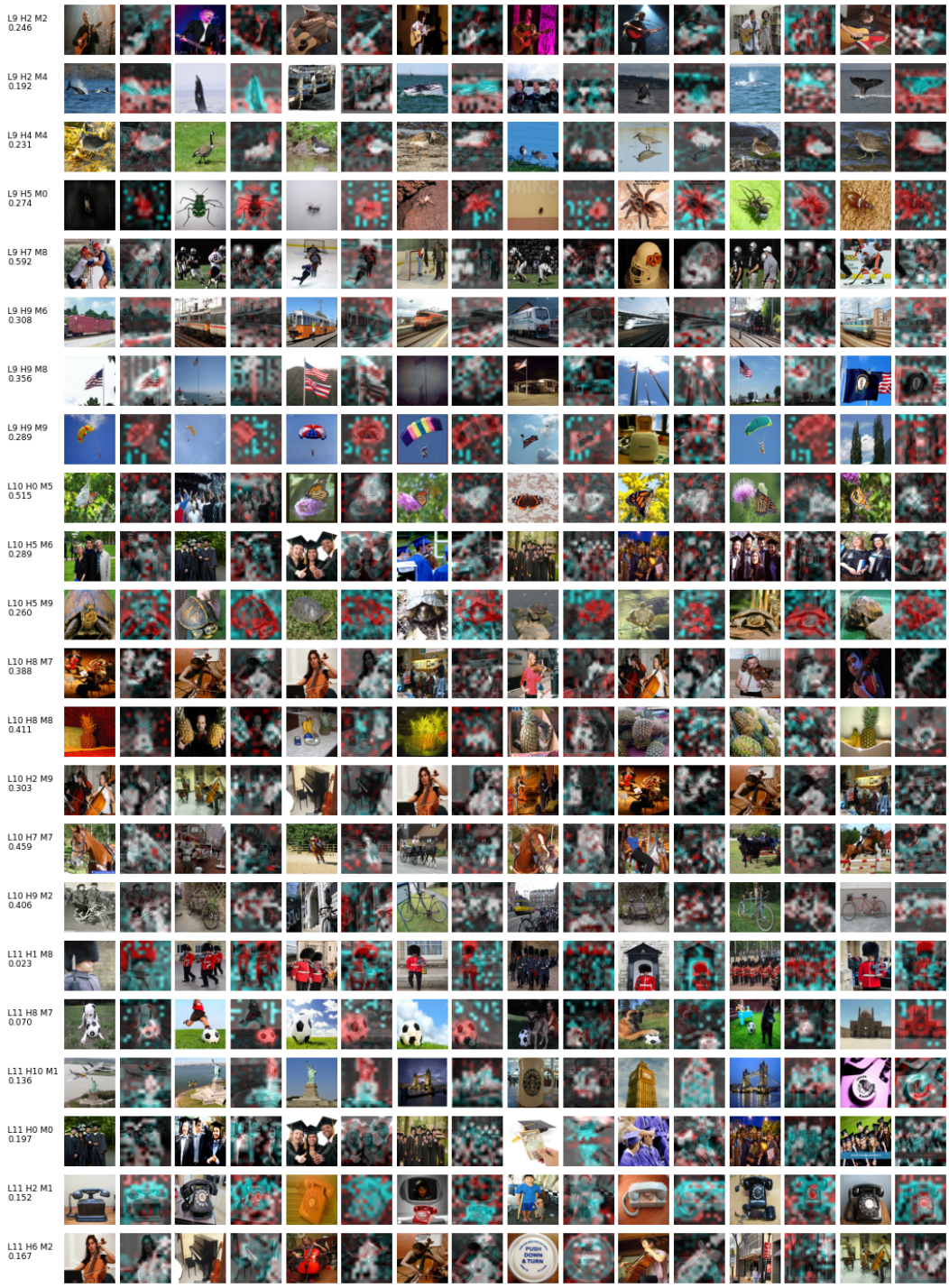

Figure S15. Examples of semantic singular modes in clip-vit-base-patch16 (part 3).