# On the Information Geometry of Vision Transformers

**Sonia Joseph** SONIA.JOSEPH@MILA.QUEBEC
*Mila and McGill University, Montréal, QC, Canada*

**Arna Ghosh** GHOSHARN@MILA.QUEBEC
*Mila and McGill University, Montréal, QC, Canada*

**Kumar Krishna Agrawal** KAGRAWAL@BERKELEY.EDU
*UC Berkeley, CA, USA*

**Blake A. Richards** BLAKE.RICHARDS@MILA.QUEBEC
*Mila and McGill University, Montréal, QC, Canada*
*Learning in Machines and Brains Program, CIFAR, Toronto, ON, Canada*

## Abstract

Understanding the structure of high-dimensional representations learned by Vision Transformers (ViTs) provides a pathway toward developing a mechanistic understanding and further improving architecture design. In this work, we leverage tools from information geometry to characterize representation quality at a per-token (*intra-token*) level as well as across pairs of tokens (*inter-token*) in ViTs pretrained for object classification. In particular, we observe that these high-dimensional tokens exhibit a characteristic spectral decay in the feature covariance matrix. By measuring the rate of this decay (denoted by $\alpha$) for each token across transformer blocks, we discover an $\alpha$ signature, indicative of a transition from lower to higher effective dimensionality. We also demonstrate that tokens can be clustered based on their $\alpha$ signature, revealing that tokens corresponding to nearby spatial patches of the original image exhibit similar $\alpha$ trajectories. Furthermore, for measuring the complexity at the sequence level, we aggregate the correlation between pairs of tokens independently at each transformer block. A higher average correlation indicates a significant overlap between token representations and lower effective complexity. Notably, we observe a U-shaped trend across the model hierarchy, suggesting that token representations are more expressive in the intermediate blocks. Our findings provide a framework for understanding information processing in ViTs while providing tools to prune/merge tokens across blocks, thereby making the architectures more efficient.

## 1. Introduction

Vision transformers (ViTs) have recently revolutionized computer vision, excelling in tasks like image classification and object detection (Dosovitskiy et al. (2020)). Instead of local convolutions, ViTs employ global self-attention, enhancing representation quality. However, understanding the information processing in ViTs at a per-token as well as sequence level remains a challenge (Raghu et al., 2021). Training ViTs is also demanding due to their computational complexity and the need for larger datasets and extended training cycles compared to CNNs (Deng et al., 2009; Sun et al., 2017; Touvron et al., 2021). While prior research has explored certain functional properties of Vision Transformers (ViTs), such as their responses to specific image transformations (Naseer et al., 2021), there has been limited investigation into understanding the information geometry of these models.

Information geometry formally characterizes the high-dimensional representation manifolds learned by neural networks and bridges the understanding between individual units and overall network behavior (Chung and Abbott (2021)). Recent studies in vision neuroscience found that mouse and macaque V1 representations demonstrate a power-law decay in their eigenspectrum (Stringer et al., 2019; Kong et al., 2022), with the decay coefficient $\alpha$ indicating the representation manifold properties like smoothness and capacity. Parallel studies in deep learning have demonstrated the significance of such power-law decay in the representation eigenspectrum to assess the network's performance (Ghosh et al., 2022; Agrawal et al., 2022) and adversarial robustness (Nassar et al., 2020).

In this paper, we characterize the ViT's information geometry within tokens (*intra-token*) by inspecting their eigenspectrum and among tokens (*inter-token*) by inspecting their correlation structure. The two-level analysis provides insight into two complementary properties of the representation manifold, namely at the token and the sequence level. We show that ViTs have a unique intra-token $\alpha$ signature, in which each token exhibits a high $\alpha$ at the initial layers and gradually exhibits lower alpha in the later layers. In other words, this indicates a transition from intrinsically low-dimensional to high-dimensional structures within the network. Second, for measuring the complexity at the sequence level, we aggregate the correlation between pairs of tokens independently at each transformer block. A higher average correlation indicates a significant overlap between token representations and lower effective complexity. Notably, we observe a U-shaped trend across the model hierarchy, suggesting that token representations are more expressive in the intermediate blocks. Our findings have implications for token pruning and merging techniques, paving way for advancements in mechanistic interpretability and designing more efficient transformer architectures.

## 2. Background

### 2.1. Sequence modeling with ViTs

The ViT architecture, inspired from NLP (Vaswani et al., 2017), divides image $x \in \mathbb{R}^{H \times W \times C}$ into patches that are then embedded as $x_p \in \mathbb{R}^{P^2 \cdot C}$, termed "token" in this paper. Transformers use self-attention blocks that process these tokens in a sequence-to-sequence framework. Namely, each block employs multi-head attention $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V$. A classification token (CLS) is optionally added to pool information across the sequence for downstream tasks. Interestingly, Raghu et al. (2021) noted that ViTs preserve more information from CLS token early in the network and switch to preserving more information from spatial tokens later in the network.

More recent work has found token representations in ViTs to be redundant, thereby leveraging this redundancy to prune (Meng et al., 2022; Yin et al., 2022; Kong et al., 2021; Rao et al., 2021) and/or merge tokens (Bolya et al., 2022; Marin et al., 2021). However, each token's representation geometry in pretrained ViTs as well as the underlying reason behind the redundancy in token representations remains poorly understood.

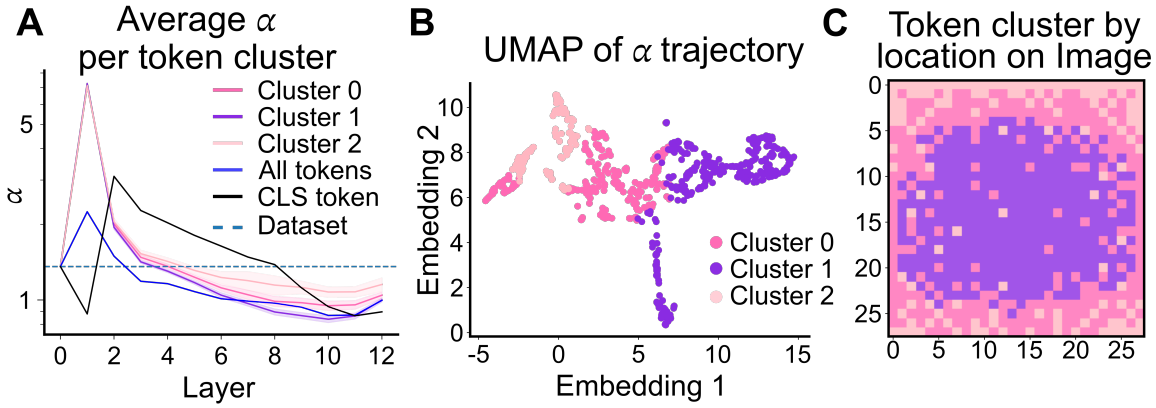



Figure 1: Intra-token analysis for ViT on CIFAR-10. We ran UMAP and k-means on the $\alpha$ trajectories for the spatial tokens and found three distinct clusters. **(A)** $\alpha$ trajectories for the three token clusters on the spatial token trajectories (pink, purple, and light pink), the CLS token (black), and all tokens together (blue). The y-axis is the $\alpha$ value on the log scale, and the x-axis is the dataset (0), and the 12 layers of the ViT (1-12). **(B)** UMAP and k-means clustering of the alpha trajectories. **(C)** The spatial location of each k-means cluster, mapping to the original image's corners, edges, and main body.

### 2.2. Eigenspectrum decay

For a parameterized function $f_\theta : X \to \mathbb{R}^D$, the empirical feature covariance matrix $\Sigma_N(f_\theta)$ provides insight into the overlap between different directions in the feature space. Formally (assuming centered features), $\Sigma_N(f_\theta) = \frac{1}{N}\sum_{i=1}^{N} f_\theta(x_i) f_\theta(x_i)^T$, where $x_1, ..., x_N \sim X$

Eigendecomposition of $\Sigma_N(f_\theta)$ reveals the variance explained by the principal components of the representation space. Using spectral decomposition, $\Sigma_N(f_\theta) = U\Lambda U^T$, where $U$ contains eigenvectors and $\Lambda$ is a diagonal matrix with nonnegative eigenvalues $\lambda_1 \geq ... \geq \lambda_m \geq 0$, where $m = \min(N, D)$ is the rank. Notably, Ghosh et al. (2022) observe that for well-trained models, this eigenspectrum is well approximated by power-law decay, where $\lambda_j \propto j^{-\alpha}$ for $\alpha > 0$. Here, $\alpha$ is the slope of the power law and coefficient of eigenvalue decay. Intuitively, a smaller $\alpha$ indicates slower decay and, therefore, higher effective rank, whereas high $\alpha$ indicates rapid decay, corresponding to intrinsically low-dimensional encodings.

## 3. Experiments

We evaluated individual token representations of a ViT pretrained on ImageNet on CIFAR-10 (Krizhevsky et al., 2009) and STL-10 (Quattoni and Torralba, 2009). To assess the representation geometry for individual tokens, we computed the representation covariance matrix and subsequently calculated $\alpha$ (Appendix B, Fig. 3). Notably, there was no significant difference in $\alpha$ between the train and test sets (Appendix D, Fig. 6), so we refer to the test set estimate unless otherwise mentioned. To assess the sequence-level complexity, we calculated token correlations by taking the Pearson's correlation between the flattened representation matrices ($\#samples \times \#feats$) for every pair of tokens. Further details are in Appendix A.

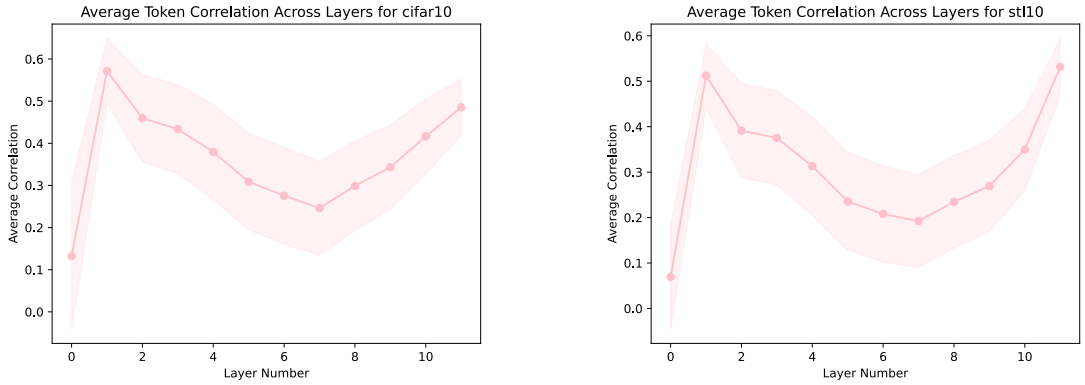



Figure 2: Inter-token analysis for ViT on CIFAR-10 & STL-10. First, we compute the correlation across flattened feature matrices for every pair of tokens (see Appendix). Next, we take the mean correlation of the resulting lower triangular matrix, thereby indicating the the expected correlation between any pair of tokens in each layer. Our results reveal a U-shaped structure, indicating that tokens are least correlated at an intermediate layer but highly correlated in the early and late layers.

### 3.1. Intra-token Analysis

For each spatial token representation, we found the $\alpha$ value starts high and then non-monotonically decreases toward $\alpha \approx 1$ (Fig.1 A). Intuitively, these tokens start lower dimensional in the first layer of the network, then non-monotonically become higher dimensional as they progress through the network. Interestingly, most spatial tokens reach their highest dimensionality towards the end of the network but before the last layers, when they slightly decrease in dimensionality. Compared to the spatial tokens, the CLS token starts higher dimensional, becomes lower dimensional around the second layer, and then non-monotonically increases in dimensionality. The decrease in dimensionality around the second layer suggests the CLS token may "take up" information from the spatial tokens in the preceding layer. Our results are in line with the findings of Ghosh et al. (2022), except they were looking at the CLS token only. We found similar results on both CIFAR-10 and STL-10 (Appendix D).

We ran UMAP on the 784 $\alpha$ trajectories of the spatial tokens (excluding the CLS-token), clustered the resulting embeddings with k-means, and found three distinct clusters (Fig.1 B) that correspond to the edges, corners, and body of the original image (Fig.1 C). The $\alpha$ trajectory of Cluster 2 (light pink), corresponding to the corners of the image, signifies lower dimensionality than the other clusters, which intuitively follows because the corners of an image usually contain less information. The clusters become more high-dimensional, according to their $\alpha$ trajectories, towards the center of the image. Our analysis suggests the $\alpha$ calculation illuminates the information density of each token, which could play an important role in token pruning and token merging.

### 3.2. Inter-token Analysis

We calculated Pearson's correlation between tokens and found that tokens are less correlated in the first layer, then rapidly become more correlated in the second layer, and finally follow a U-shape for the rest of the network, ending with a high correlation (Fig. 2). See Appendix C for layer-by-layer, inter-token correlation plots on CIFAR-10 and STL-10. It remains unclear whether increasing token correlations in the network's later layers is functional or redundant, suggesting potential improvements in transformer architecture.

## 4. Conclusion

For the first time, we explored ViT's information geometry at both intra-token and inter-token levels, revealing a unique $\alpha$ trajectory converging to $\alpha \approx 1$, indicating token dimensionality changes across layers. $\alpha$ trajectories and token correlations suggest pruning and merging potential, emphasizing information geometry's role in transformer optimization. Our findings have implications for computer vision, transformer efficiency, model interpretability, and mechanistic interpretability.

## Acknowledgements

This research was generously supported by the Bank of Montreal (S.J.); Vanier Canada Graduate scholarship (A.G.); NSERC (Discovery Grant: RGPIN-2020-05105; Discovery Accelerator Supplement: RGPAS-2020-00031; Arthur B. McDonald Fellowship: 566355-2022) and CIFAR (Canada AI Chair; Learning in Machine and Brains Fellowship). This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada. The authors acknowledge the material support of NVIDIA in the form of computational resources.

## References

Kumar K Agrawal, Arnab Kumar Mondal, Arna Ghosh, and Blake Richards. $\alpha$-req: Assessing representation quality in self-supervised learning by measuring eigenspectrum decay. *Advances in Neural Information Processing Systems*, 35:17626–17638, 2022.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

SueYeon Chung and L.F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70: 137–144, 2021. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2021.10.010. URL https://www.sciencedirect.com/science/article/pii/S0959438821001227. Computational Neuroscience.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Arna Ghosh, Arnab Kumar Mondal, Kumar Krishna Agrawal, and Blake Richards. Investigating power laws in deep representation learning. *arXiv preprint arXiv:2202.05808*, 2022.

Nathan CL Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022.

Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Mengshu Sun, Wei Niu, Xuan Shen, Geng Yuan, Bin Ren, Minghai Qin, et al. Spvit: Enabling faster vision transformers via soft token pruning. *arXiv preprint arXiv:2112.13890*, 2021.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860*, 2021.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022.

Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

Josue Nassar, Piotr Sokol, SueYeon Chung, Kenneth D Harris, and Il Memming Park. On 1/n neural representation and robustness. *Advances in Neural Information Processing Systems*, 33:6211–6222, 2020.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, pages 413–420. IEEE, 2009.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571 (7765):361–365, 2019.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, abs/1707.02968, 2017. URL http://arxiv.org/abs/1707.02968.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10809–10818, 2022.

## Appendix A. Additional Details on Experimental Setup

For the intra-token analysis, we evaluated individual token activations in the intermediate layers of a ViT pretrained on ImageNet, taken from the timm library (Wightman (2019)). The ViT divides 3 x 224 x 224 images from CIFAR-10 (Krizhevsky et al. (2009)) and STL-10 (Quattoni and Torralba (2009)) into a patch size of 8 for a total of 785 tokens including the CLS token. The images were fed into the ViT with batch size 128. For each token activation for each layer, we evaluated $\alpha$ across all data points by calculating the covariance matrix and fitting a power law to the eigenspectrum (Appendix B, Fig. 3). We also calculated the $\alpha$ trajectory of the concatenated token representations (including the CLS token). For the inter-token analysis, we calculated token correlations by taking the Pearson's correlation between every pair of tokens across all data points in the test set.

## Appendix B. Eigenspectrum fits

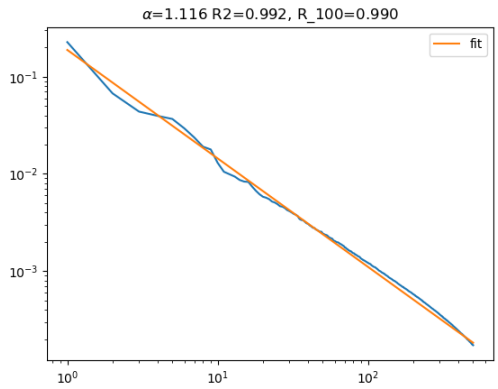



Figure 3: Example fit of alpha on the eigenspectrum for token 1 at layer 217 on CIFAR-10. The blue line is the empirical measurement while the orange line is the calculated fit. For more details, see Sections 2.2 and 3.

# Appendix C. Token Correlations

## C.1. CIFAR-10

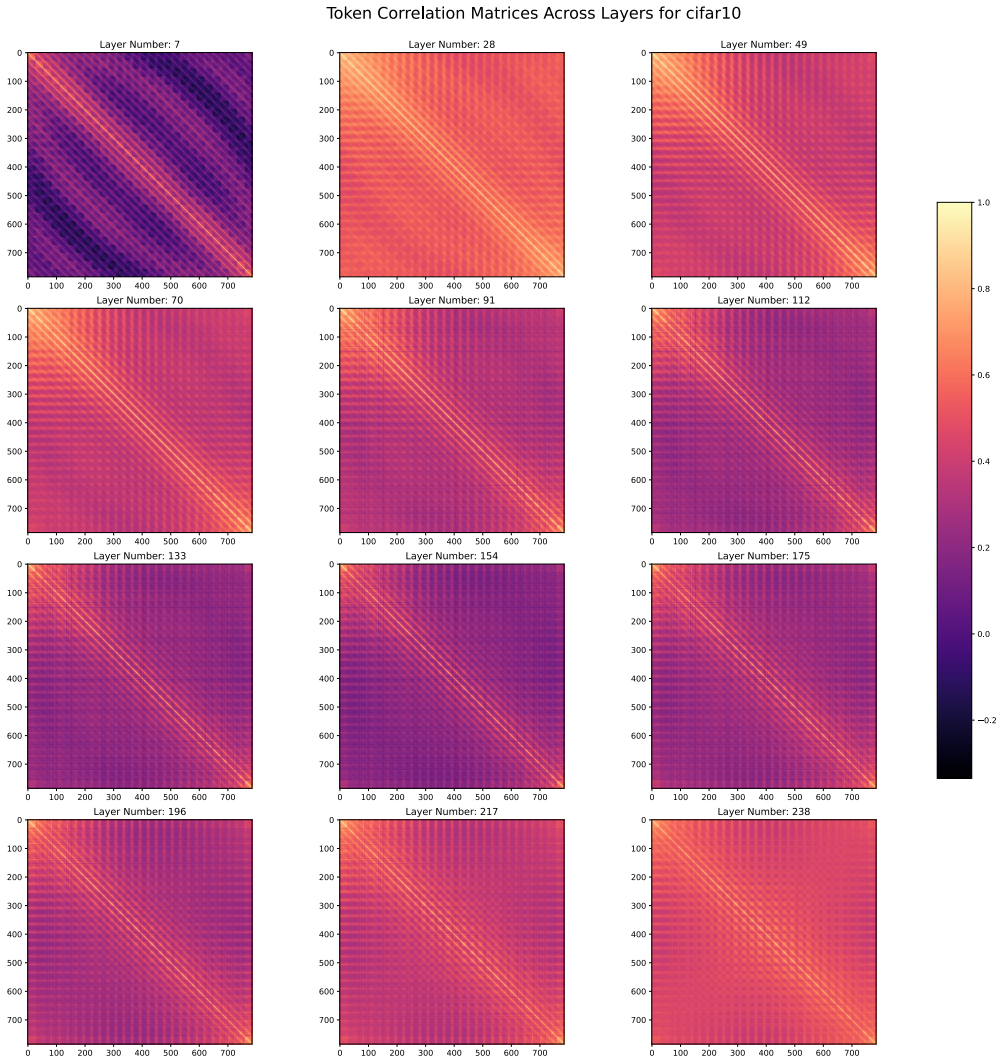



Figure 4: Token activation correlations per layer for ViT on CIFAR-10. Each matrix represents a layer of 785 tokens x 785 tokens, for a total of 12 layers. For more details, see Section 3.2.

## C.2. STL-10

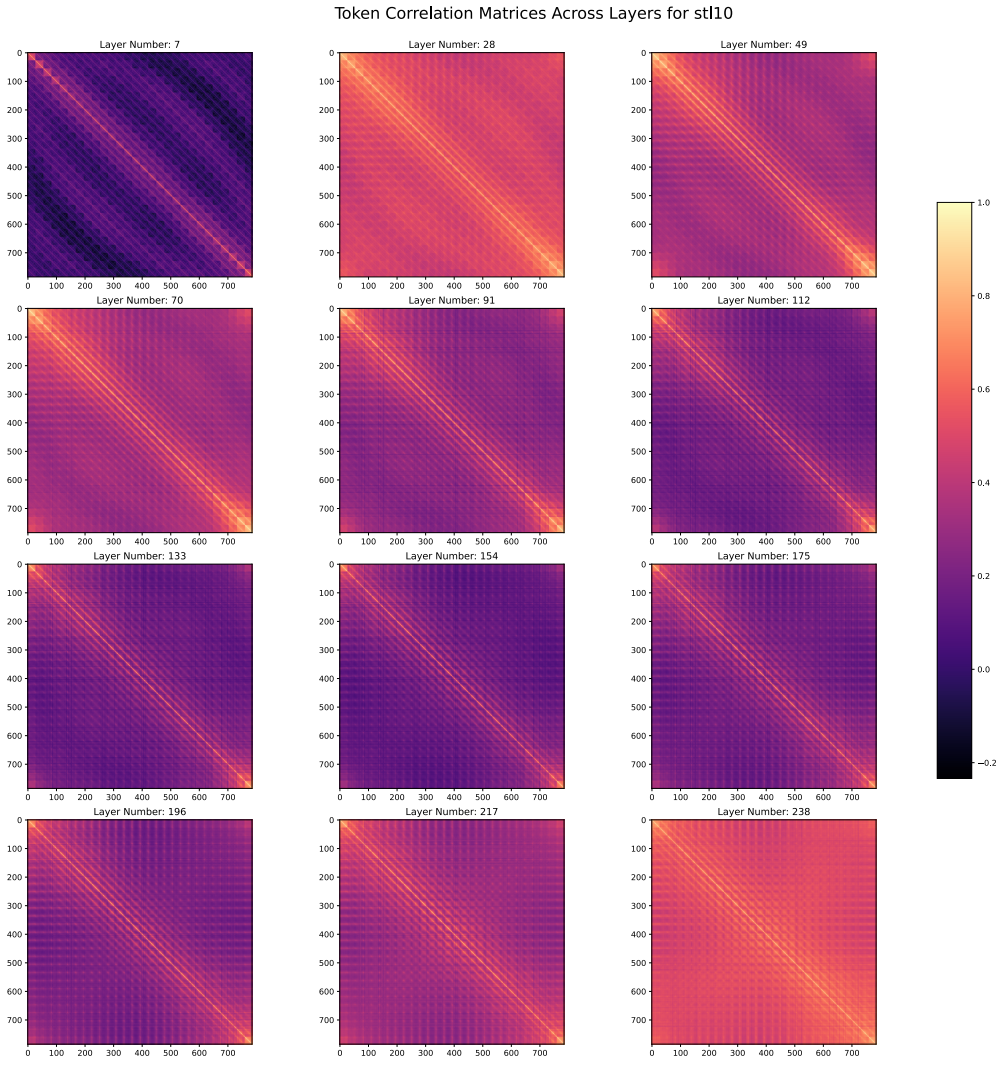



Figure 5: Token activation correlations per layer for ViT on STL-10 test set. For more details, see Section 3.2.

## Appendix D. More Alpha Results

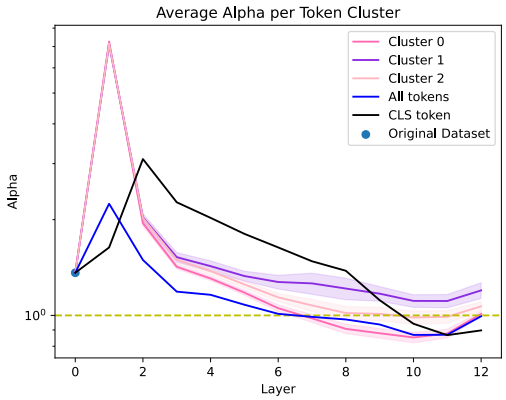
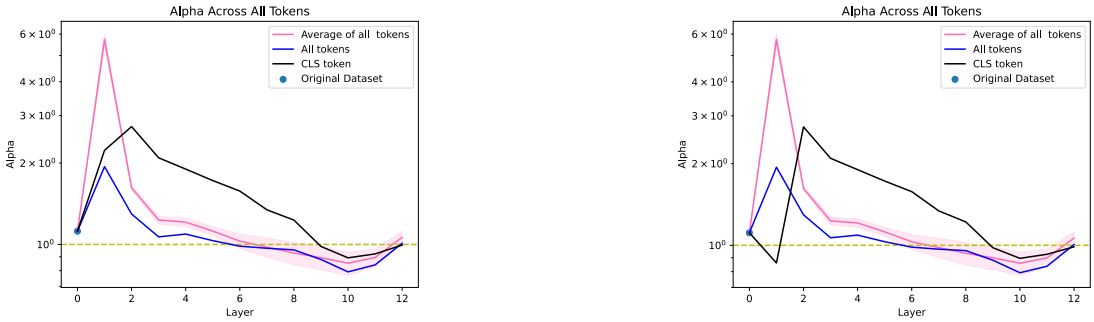



Figure 6: Alpha trajectories for ViT on CIFAR-10 train set. The $\alpha$ trajectory is very similar to that on the CIFAR-10 test set, so we refer to the empircal $\alpha$ calculation throughout the paper as the calculation on the test set. See Sections 3 and 3.1 for more details.


(a) (b)

Figure 7: K-means on ViT alpha trajectories for the CIFAR-10 train set. (a) UMAP of K-means clustering on the alpha trajectories. (b) The corresponding spatial location of each cluster on the original image. The clusters map to the corners, edges, and main body of the image. The main body of the paper used the results from the CIFAR-10 test set. For more details, see Section 3.1.




(a) (b)

Figure 8: The alpha trajectories on STL-10 for the CLS token (black), the average of all tokens (pink), and all tokens together (blue) on the (a) train and (b) test datasets. For more details, see Section 3.1.