FTC: Fourier Transcoder for Monosemantic Sparse Feature Decomposition in Vision Language Model

Kyeongjun Cho [1] **Juheon Kang** [1] **Hyunseo Shin** [1] **Hongseok Oh** [1] **Wonseok Hwang** [1]

Abstract

Sparse autoencoders (SAEs) and transcoders (TCs) have been successfully applied in interpreting neural activations in large language models (LLMs), particularly those in MLP sublayers, through sparse feature representations. Building on this success, SAEs and TCs have also been extended to vision-language models (VLMs), including CLIP, with promising results. However, significant challenges remain, including feature polysemanticity and incomplete disentanglement. In this work, we introduce the Fourier Transcoder (FTC), a transcoder architecture based on a Fourier basis, to analyze MLP sublayers in a VLM. We show that FTC discovers more monosemantic features and reduces the number of dead features in CLIP compared to TopK TC.

1. Introduction

Approaches from mechanistic interpretability research, such as sparse autoencoders (SAEs) and transcoders (TCs) (Bricken et al., 2023; Templeton et al., 2024; Dunefsky et al., 2024), have proven effective for interpreting LLMs (Huben et al., 2024) and are increasingly being explored in VLMs (Pach et al., 2026; Shen et al., 2026). Particularly, TCs offer a useful framework for circuit-level analysis of multimodal features within MLP sublayers (Yang et al., 2026). However, the faithfulness of circuit analyses depends on learning sparse and monosemantic features, which remains challenging in practice due to feature polysemanticity and incomplete disentanglement.

Recent work has shown that spectral perspectives can provide useful insight into how neural networks represent and process information, and that structured Fourier feature mappings can improve how complex signals are parameterized in neural models (Zhou et al., 2024; Rahaman et al., 2019; Tancik et al., 2020). These observations suggest that a Fourier-structured basis may offer a useful inductive bias for learning more disentangled latent representations.

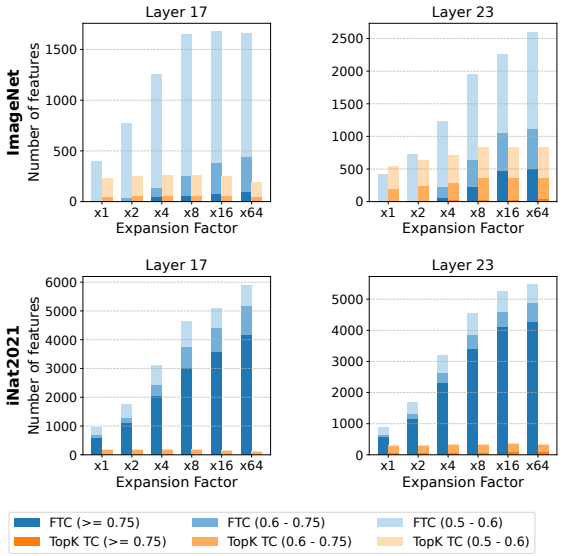

Figure 1. Number of features in three monosemanticity score ranges (0.5–0.6, 0.6–0.75, and > 0.75) for FTC (blue) and TopK TC (orange) across expansion factors, evaluated on ImageNet (up) and iNat2021 (down) at layers 17 and 23. FTC generally yields more features that show high monosemanticity score compared to TopK TC in these settings, with a larger advantage on iNat2021.

In this context, we introduce the Fourier Transcoder (FTC), a TC that parameterizes its encoder using a Fourier basis to learn sparse latent features. Our approach represents activations as combinations of coefficients over a structured basis, with the goal of improving sparse feature decomposition.

Our main contributions are as follows:

- We introduce FTC, a Fourier-basis-based transcoder that learns sparse feature representations in an overcomplete Fourier coefficient space.
- Experiments on CLIP show that FTC discovers more monosemantic features and fewer dead features that are never activated on the validation set, than TopK TC across multiple dictionary sizes.

[1]Department of AI, University of Seoul, Seoul, South Korea. Correspondence to: Wonseok Hwang <wonseok.hwang@uos.ac.kr>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

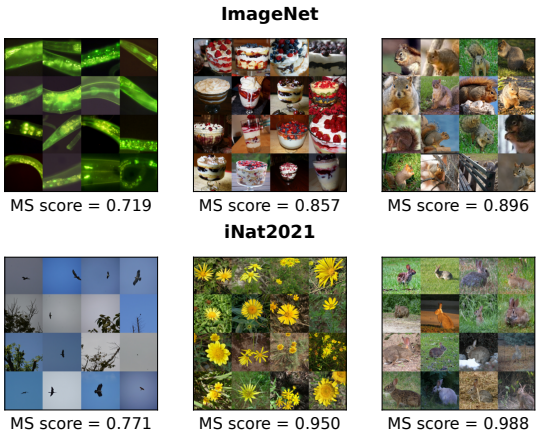

Figure 2. For each representative FTC feature, we collect the 16 images with the highest activation values over the full image set and visualize them. The top row shows features learned from ImageNet, and the bottom row shows features learned from iNat2021. These high-MS features correspond to visually coherent and human-interpretable concepts.

2. Preliminaries

2.1. Transcoders

The goal of TCs is to approximate the computation of an MLP sublayer as a sparse combination of interpretable features. TCs consist of an encoder layer and a decoder layer. Given an input activation $x \in \mathbb{R}^{d_{\text{model}}}$ from an MLP sublayer, the encoder maps x to sparse latent features, while the decoder reconstructs the corresponding MLP output from these sparse features. Formally:

$$z_{\text{TC}}(x) = \text{ReLU}(W_{\text{enc}}x + b_{\text{enc}}),$$
$$\text{TC}(x) = W_{\text{dec}} z_{\text{TC}}(x) + b_{\text{dec}},$$

where $W_{\text{enc}} \in \mathbb{R}^{d_{\text{features}} \times d_{\text{model}}}$, $b_{\text{enc}} \in \mathbb{R}^{d_{\text{features}}}$, $W_{\text{dec}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{features}}}$, $b_{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$ are the learnable parameters. Here, d_{features} denotes the latent dimension of the TCs, corresponding to the number of learned features, and d_{model} denotes the dimensionality of the MLP input activation. A widely used variant is the TopK TC, which enforces sparsity in the latent space by retaining only the top-k activations of $z_{\text{TC}}(x)$ before decoding. Its direct control over sparsity through k makes it practical to tune and effective at balancing reconstruction quality against sparsity (Gao et al., 2025).

2.2. Fourier Transform

The Fourier transform decomposes an input signal into a sum of periodic components with different frequencies. Formally, for a function $f(x)$, the Fourier transform is defined as:

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)\, e^{-i\omega x}\, dx,$$

where x and ω denote the original-domain and frequency variables, respectively, and $\hat{f}$ is the transform of f.

Because the activations considered in this work are real-valued, we restrict attention to the real-valued Fourier basis. In addition, since our analysis is performed over finite-dimensional activation vectors, we use the discrete cosine transform (DCT), which provides a real-valued decomposition of discrete inputs. In particular, the DCT-II of a discrete input x_n of length N is given by (Strang, 1999):

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \tfrac{1}{2}\right) k\right].$$

where $n = 0, \ldots, N-1$ indexes the input coordinate, $k = 0, \ldots, N-1$ indexes the cosine frequency component, x_n denotes the input value at index n, and X_k denotes the corresponding coefficient in the cosine basis.

3. Fourier Transcoder

We introduce the Fourier Transcoder (FTC), a variant of the TopK TC that employs a learnable Fourier basis in the encoder. Unlike TopK TC, which performs sparse feature selection in a learned latent space, FTC performs sparse feature selection in a Fourier coefficient space.

3.1. Architecture

For an input activation $x \in \mathbb{R}^{d_{\text{model}}}$, the encoder of FTC can be written as follows:

$$z_{\text{FTC}}(x) = W_{\text{enc}}x + b_{\text{enc}},$$
$$\tilde{z}_{\text{FTC}}(x) = B_{\omega}^{\top} z_{\text{FTC}}(x),$$

The input x is first mapped by $W_{\text{enc}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ to an intermediate vector $z_{\text{FTC}}(x) \in \mathbb{R}^{d_{\text{model}}}$. The learnable Fourier basis $B_{\omega} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{features}}}$ consists of d_{features} Fourier bases having dimension d_{model}. Specifically, the (n, j)-th entry of B_{ω} is defined as:

$$(B_{\omega})_{n,j} = \cos\left(\frac{\pi\left(n + \frac{1}{2}\right)\omega_j}{d_{\text{model}}}\right),$$

where j is the index for each basis and n is the index for an element within an individual basis. Here only ω_j (frequencies) are learnable parameters.

Thus B_{ω}^{T} projects $z_{\text{FTC}}(x)$ onto these bases and therefore each entry of $\tilde{z}_{\text{FTC}}(x)$ is the coefficient associated with one basis element, and corresponds to one latent feature.

To construct an overcomplete dictionary, we set

$$d_{\text{features}} = d_{\text{model}} \cdot \varepsilon,$$

where ε denotes the expansion factor. Rather than restricting the dictionary to the standard d_{model} DCT frequencies, we

Table 1. R^2 of FTC and TopK TC on ImageNet and iNat2021.

Dataset	Layer	FTC						TopK TC					
		×1	×2	×4	×8	×16	×64	×1	×2	×4	×8	×16	×64
ImageNet	11	0.715	0.722	0.723	0.720	0.720	0.718	0.733	0.735	0.739	0.735	0.735	0.727
	17	0.587	0.605	0.608	0.607	0.599	0.326	0.622	0.626	0.631	0.628	0.623	0.599
	22	0.593	0.610	0.614	0.613	0.610	0.607	0.707	0.734	0.745	0.750	0.750	0.742
	23	0.691	0.698	0.698	0.696	0.695	0.690	0.739	0.750	0.757	0.762	0.763	0.760
	last	0.730	0.717	0.705	0.698	0.698	0.692	0.679	0.701	0.720	0.737	0.752	0.773
iNat2021	11	0.742	0.748	0.750	0.749	0.749	0.743	0.759	0.759	0.759	0.757	0.754	0.743
	17	0.707	0.709	0.718	0.714	0.709	0.683	0.720	0.718	0.711	0.709	0.701	0.678
	22	0.734	0.740	0.741	0.733	0.732	0.728	0.759	0.766	0.769	0.769	0.773	0.766
	23	0.788	0.790	0.784	0.785	0.783	0.782	0.803	0.803	0.808	0.808	0.811	0.805
	last	0.791	0.779	0.765	0.756	0.768	0.767	0.765	0.770	0.780	0.784	0.788	0.794

initialize the learnable frequencies with uniform intervals,

$$\omega_j^{(0)} = \frac{j-1}{\varepsilon}, \qquad j = 1, \ldots, d_{\text{features}}.$$

After projection onto the Fourier basis, FTC enforces sparsity in the Fourier coefficient space by applying TopK activation to the projected representation:

$$\hat{z}_{\text{FTC}}(x) = \text{TopK}(\tilde{z}_{\text{FTC}}(x), k).$$

The sparse Fourier coefficients are then mapped back to the original activation space via B_ω and decoded to reconstruct the MLP output:

$$z_{\text{dec}}(x) = B_\omega \hat{z}_{\text{FTC}}(x),$$
$$\text{FTC}(x) = W_{\text{dec}} z_{\text{dec}}(x) + b_{\text{dec}}.$$

where $W_{\text{dec}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ and $b_{\text{dec}} \in \mathbb{R}^{d_{\text{model}}}$ are learnable decoder parameters. Here, z_{dec} is a linear combination of surviving top-K features (coefficients of the Fourier basis).

Because only the frequency vector ω is learned in the Fourier basis, and because W_{enc} and W_{dec} operate in d_{model}-dimensional space, FTC requires fewer learnable parameters than TopK TC (Appendix A).

Training Objective The training objective of FTC is defined as follows:

$$\mathcal{L}_{\text{FTC}} = \|\text{MLP}(x) - \text{FTC}(x)\|_2^2.$$

As with standard transcoders, FTC is trained to approximate the output of the MLP sublayer by using the original MLP output as the reconstruction target.

4. Experiments

4.1. Experimental Setup

We evaluate FTC against standard TopK TC on the MLP sublayers of pretrained CLIP ViT-L/14-336px (Radford et al., 2021). Following prior work on SAE-based analysis of CLIP, our experiments focus on layers $\{11, 17, 22, 23\}$ and the output of the final projection layer (Pach et al., 2026). We use images from ImageNet (Deng et al., 2009) and iNat2021 (Van Horn et al., 2021) across a range of dictionary sizes: {1024, 2048, 4096, 8192, 16384, and 65536}. All experiments were conducted on three NVIDIA RTX A6000 GPUs.

4.2. Metrics

We evaluate learned features using monosemanticity (MS) score for semantic quality and R^2 for reconstruction quality.

Monosemanticity Score We use monosemanticity (MS) score (Pach et al., 2026), which measures the semantic coherence, in an embedding space, of image pairs that activate the same feature.

Given a set of images $I = \{x_n\}_{n=1}^N$ and a pretrained image encoder E, we first compute the pairwise similarity matrix S, where

$$S_{nm} := \frac{E(x_n) \cdot E(x_m)}{\|E(x_n)\| \, \|E(x_m)\|}.$$

A feature k is the k-th entry of the sparse representation: $z_{\text{TC}}(x_n)$ for TopK TC and $\hat{z}_{\text{FTC}}(x_n)$ for FTC. Let a_n^k denote the activation of feature k on image x_n and $\tilde{a}_n^k$ be the corresponding min-max normalized activation:

$$\tilde{a}_n^k = \frac{a_n^k - \min_{n'} a_{n'}^k}{\max_{n'} a_{n'}^k - \min_{n'} a_{n'}^k}.$$

Then the relevance weight for an image pair is defined as

$$r_{nm}^k := \tilde{a}_n^k \tilde{a}_m^k.$$

Finally, the monosemanticity score of feature k is given by

$$\text{MS}_k := \frac{\sum_{1 \le n < m \le N} r_{nm}^k \, s_{nm}}{\sum_{1 \le n < m \le N} r_{nm}^k}.$$

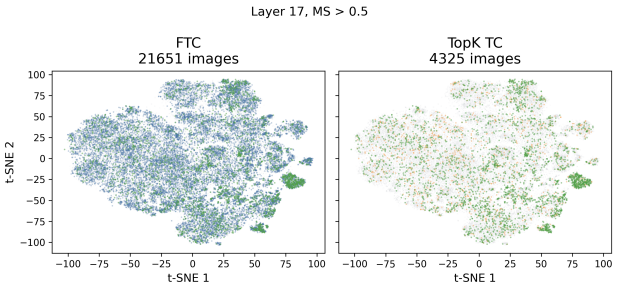

Figure 3. Union coverage of high-MS features on the global t-SNE embedding of the iNat validation set. For Layer 17 with expansion 16, each panel shows the union of top-30 activated images covered by FTC or TopK TC, with images covered by both models highlighted in green.

Thus, for each layer and expansion factor, we obtain one MS score for each feature. The MS score takes values in the range $[-1, 1]$. A higher MS score indicates that the activated images are more similar in the embedding space, suggesting a more narrowly focused and thus more monosemantic feature. A lower MS score indicates activation over a more diverse set of images, suggesting a less monosemantic and potentially more polysemantic feature.

Reconstruction Quality We evaluate the reconstruction performance of the TCs using $R^2 = 1 - \mathrm{FVU}$, which measures how well the reconstructed output $\hat{y}$ explains the variance of the target activation y. Here FVU indicates the fraction of variance unexplained (FVU) and is given by

$$\mathrm{FVU} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\mathrm{MSE}}{\mathrm{Var}(y)}.$$

where y_i and $\hat{y}_i$ denote the target and reconstructed MLP outputs, respectively, $\bar{y}$ is the mean target activation, and n is the number of samples. Higher R^2 indicates more accurate reconstruction of the original MLP output.

5. Results

5.1. Evaluating the Monosemanticity of Learned Features

To test FTC, we analyze CLIP on two image classification tasks–ImageNet and iNat2021. The results show that FTC achieves higher MS scores than standard TopK TC across most settings (Tables 2, 3 in Appendix). In particular, FTC consistently attains higher maximum MS scores and substantially fewer dead features (Figures 5, 6 in Appendix). Additional quantitative results are reported in Appendices B and C.

To quantify high-MS features, Figure 1 groups them into three MS ranges: 0.5–0.6, 0.6–0.75, and > 0.75. On ImageNet, FTC yields more high-MS features than TopK TC at layers 11 and 17, while the two become comparable in later layers (Figures 7, 8 in Appendix). On iNat2021, FTC consistently outperforms TopK TC across all evaluated layers, with the gap especially large at expansion factor 64, where FTC produces over three times more features with MS $>$ 0.5. These results suggest that FTC improves both the peak monosemantic quality of learned features and the total number of high-MS features, with the advantage becoming more pronounced on the more fine-grained iNat2021.

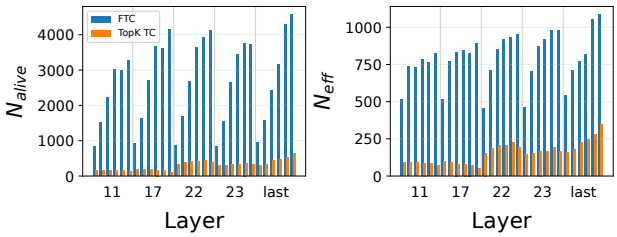

Figure 4. Results on iNat2021 for the number of alive features (N_{alive}, left) and the effective number of features (N_{eff}, right). Although N_{eff} is reduced relative to N_{alive}, FTC remains higher than TopK TC, indicating that its advantage persists even after accounting for feature redundancy.

To assess whether more high-MS features correspond to more meaningful feature discovery, we projected the union of top-30 activated images onto the global t-SNE space of iNat validation image embeddings for FTC and TopK TC. Figure 3 shows that FTC covers the image space more broadly than TopK TC, as detailed in Appendix F.

5.2. Effective Feature Rank Analysis

Figure 2 visualizes the top 16 activating images for selected high-MS features, showing that they often capture visually coherent and human-interpretable concepts.

Further qualitative analysis shows that visually similar FTC feature groups often recur across nearby basis frequencies. This suggests that some learned features may be partially redundant, with nearby basis frequencies ω associated with similar sets of images. This raises the question of whether a larger number of high-MS features truly reflects a larger number of distinct and meaningful features.

To further examine this issue, we define N_{alive} as the number of features whose MS score exceeds 0.5, and draw the Pearson correlation matrix C computed from the activations of these alive features. We then quantify the effective number of independent features using the participation-ratio effective rank (Roy & Vetterli, 2007),

$$N_{\text{eff}} = \frac{\left(\sum_i C_{ii}\right)^2}{\sum_{i,j} C_{ij}^2}.$$

Figure 4 shows the results for N_{alive} and N_{eff} on iNat2021. Although the effective number of features decreases relative to N_{alive}, FTC still achieves a substantially higher N_{eff}

than standard TopK TC. On ImageNet, FTC likewise yields higher N_{eff} in the 11 and 17 layers, whereas FTC and TopK TC become largely comparable in the later layers; detailed results are provided in Appendix E. These results indicate that FTC retains a larger number of effective high-quality features even after accounting for redundancy, with a particularly clear advantage on iNat2021.

5.3. Reconstruction Quality

Table 1 shows that FTC exhibits slightly lower R^2 than standard TopK TC in some settings, particularly in later layers and at higher expansion factors, while remaining largely comparable in reconstruction quality across most configurations. These results indicate that FTC retains competitive reconstruction performance while using fewer learnable parameters than standard TopK TC.

6. Conclusion

We introduced the Fourier Transcoder (FTC), a transcoder architecture that improves monosemantic feature learning for VLM interpretation tasks. On CLIP, FTC produced more high-MS features and substantially fewer dead features than standard TopK TC. Although some partially redundant features were observed, effective-rank analysis showed that FTC still maintained a larger number of effective monosemantic features overall and FTC also achieved reconstruction quality comparable to standard TopK TC. These results suggest that introducing a learnable Fourier basis into the encoder can improve sparse feature decomposition in a VLM.

7. Acknowledgments

We thank Eunkyung Choi and Seungwon Woo for their helpful discussion during the early stages of this research. We thank the members of the IDA.L laboratory for their discussions throughout the project. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00341654).

References

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J6zHcScAo0.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Pach, M., Karthik, S., Bouniot, Q., Belongie, S., and Akata, Z. Sparse autoencoders learn monosemantic features in vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=DaNnkQJSQf.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., and Courville, A. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Roy, O. and Vetterli, M. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pp. 606–610, 2007.

Shen, S., Sun, J., Huang, Q., and Wang, S. VL-SAE: Interpreting and enhancing vision-language alignment with a unified concept set. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=1Sb0363f2y.

Strang, G. The discrete cosine transform. *SIAM review*, 41 (1):135–147, 1999.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and MacAodha, O. Benchmarking Representation Learning for Natural World Image Collections . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12879–12888, Los Alamitos, CA, USA, June 2021. IEEE Computer Society. doi: 10.1109/CVPR46437.2021.01269. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01269.

Yang, J., Xiong, T., Qian, S., Nahrstedt, K., and Wu, M. Circuit tracing in vision-language models: Understanding the internal mechanisms of multimodal thinking. *arXiv preprint arXiv:2602.20330*, 2026.

Zhou, T., Fu, D., Sharan, V., and Jia, R. Pre-trained large language models use fourier features to compute addition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=i4MutM2TZb.

A. Parameter Efficiency

We compare the number of learnable parameters of FTC with that of a standard TopK TC. For a TopK TC with input dimension d_{model}, latent dimension d_{features}, and output dimension d_{out}, the learnable parameter count is

$$P_{\text{TopK}} = d_{\text{features}} d_{\text{model}} + d_{\text{out}} d_{\text{features}} + d_{\text{features}} + d_{\text{model}} + d_{\text{out}},$$

which is dominated by the dense encoder and decoder and therefore scales as

$$P_{\text{TopK}} = \Theta\big(d_{\text{features}}(d_{\text{model}} + d_{\text{out}})\big).$$

In FTC, the learned dense feature matrices are replaced by a learned mixing matrix, a learnable frequency vector, and an output decoder. The corresponding parameter count is

$$P_{\text{FTC}} = d_{\text{model}}^2 + d_{\text{features}} + d_{\text{out}} d_{\text{model}} + d_{\text{model}} + d_{\text{out}},$$

which scales as

$$P_{\text{FTC}} = \Theta\big(d_{\text{model}}^2 + d_{\text{out}} d_{\text{model}} + d_{\text{features}}\big).$$

Under the common setting $d_{\text{out}} = d_{\text{model}}$ and $d_{\text{features}} = \varepsilon d_{\text{model}}$, the parameter counts simplify to

$$P_{\text{TopK}} = 2\varepsilon d_{\text{model}}^2 + \varepsilon d_{\text{model}} + 2d_{\text{model}},$$
$$P_{\text{FTC}} = 2d_{\text{model}}^2 + \varepsilon d_{\text{model}} + 2d_{\text{model}}.$$

Therefore, in the overcomplete regime $\varepsilon > 1$, the dominant terms yield

$$\frac{P_{\text{FTC}}}{P_{\text{TopK}}} \approx \frac{2d_{\text{model}}^2}{2\varepsilon d_{\text{model}}^2} = \frac{1}{\varepsilon},$$

showing that FTC becomes increasingly parameter-efficient relative to a standard TopK TC as the expansion factor grows.

B. Maximum/Minimum Monosemanticity Scores

In this appendix, we report the maximum and minimum monosemanticity (MS) scores on ImageNet and iNat2021 across layers and expansion factors. On ImageNet, the maximum MS scores of FTC are generally higher than those of TopK TC, although TopK TC achieves comparable or slightly better scores under some settings.

Table 2. Highest and lowest monosemanticity scores of features in the CLIP ViT-L model for the ImageNet experiment. Values are reported in the format Max / Min.

TC type	Layer	Expansion factor					
		x1	x2	x4	x8	x16	x64
TopK TC	11	0.91 / 0.46	0.61 / 0.44	0.60 / 0.45	0.83 / 0.44	0.61 / 0.43	1.00 / 0.44
	17	0.81 / 0.43	1.00 / 0.39	0.83 / 0.39	0.89 / 0.29	0.88 / 0.37	0.86 / 0.42
	22	0.83 / 0.40	0.86 / 0.34	0.90 / 0.38	0.87 / 0.34	0.88 / 0.34	0.90 / 0.33
	23	0.79 / 0.34	0.80 / 0.34	0.87 / 0.34	0.93 / 0.33	1.00 / 0.33	1.00 / 0.33
	last	0.75 / 0.41	0.76 / 0.41	0.78 / 0.41	0.79 / 0.41	0.81 / 0.41	0.84 / 0.41
FTC	11	0.72 / 0.27	0.85 / 0.37	0.94 / 0.32	0.92 / 0.25	0.89 / 0.24	0.95 / 0.22
	17	0.80 / 0.30	0.87 / 0.26	0.89 / 0.15	0.92 / 0.23	1.00 / 0.31	0.91 / 0.28
	22	0.72 / 0.33	0.86 / 0.40	0.90 / 0.28	0.93 / 0.20	0.93 / 0.29	0.94 / 0.22
	23	0.69 / 0.30	0.92 / 0.31	0.92 / 0.26	0.95 / 0.26	1.00 / 0.32	0.94 / 0.29
	last	0.56 / 0.41	0.83 / 0.40	0.89 / 0.36	0.97 / 0.28	0.95 / 0.30	1.00 / 0.23

Table 3. Highest and lowest monosemanticity scores of features in the CLIP ViT-L model for the iNat2021 experiment. Values are reported in the format Max / Min.

TC type	Layer	Expansion factor					
		x1	x2	x4	x8	x16	x64
TopK TC	11	0.74 / 0.62	0.76 / 0.62	0.77 / 0.61	0.93 / 0.45	0.92 / 0.53	0.84 / 0.61
	17	0.80 / 0.61	0.87 / 0.61	0.80 / 0.63	0.86 / 0.58	0.79 / 0.61	0.88 / 0.61
	22	0.85 / 0.61	0.86 / 0.59	0.86 / 0.60	0.97 / 0.59	0.85 / 0.48	0.91 / 0.53
	23	0.84 / 0.60	0.88 / 0.58	0.86 / 0.58	0.84 / 0.54	0.90 / 0.55	0.88 / 0.57
	last	0.82 / 0.61	0.82 / 0.61	0.82 / 0.61	0.84 / 0.61	0.84 / 0.57	0.85 / 0.57
FTC	11	1.00 / 0.49	1.00 / 0.50	1.00 / 0.49	1.00 / 0.47	1.00 / 0.42	1.00 / 0.43
	17	1.00 / 0.45	1.00 / 0.48	1.00 / 0.43	1.00 / 0.42	1.00 / 0.44	1.00 / 0.37
	22	1.00 / 0.50	1.00 / 0.50	1.00 / 0.43	1.00 / 0.45	1.00 / 0.46	1.00 / 0.41
	23	1.00 / 0.50	1.00 / 0.50	1.00 / 0.45	1.00 / 0.47	1.00 / 0.39	1.00 / 0.48
	last	1.00 / 0.49	1.00 / 0.44	1.00 / 0.49	1.00 / 0.39	1.00 / 0.45	1.00 / 0.50

C. Analysis of Dead Features

We extend the analysis in Figure 1 by additionally visualizing the number of dead features and the number of alive features above a given MS threshold. Across experiments on both ImageNet and iNat2021, TopK TC exhibits a substantially larger number of dead features than FTC. These results further suggest that FTC not only yields more high-quality features, but also makes more effective use of the feature dictionary by reducing the number of inactive features.

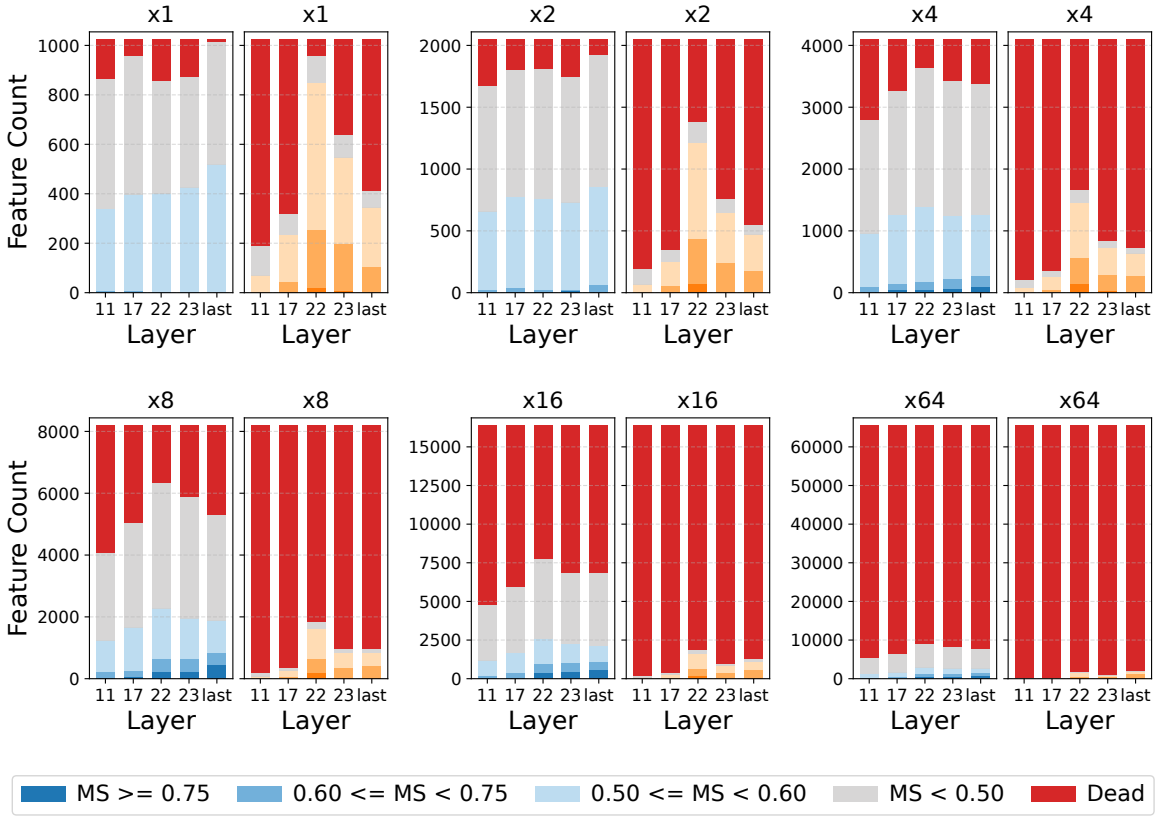

Figure 5. Comparison of the number of features on ImageNet across five categories—dead features, MS < 0.5, 0.5–0.6, 0.6–0.75, and MS > 0.75—for FTC (blue) and TopK TC (orange) across expansion factors.

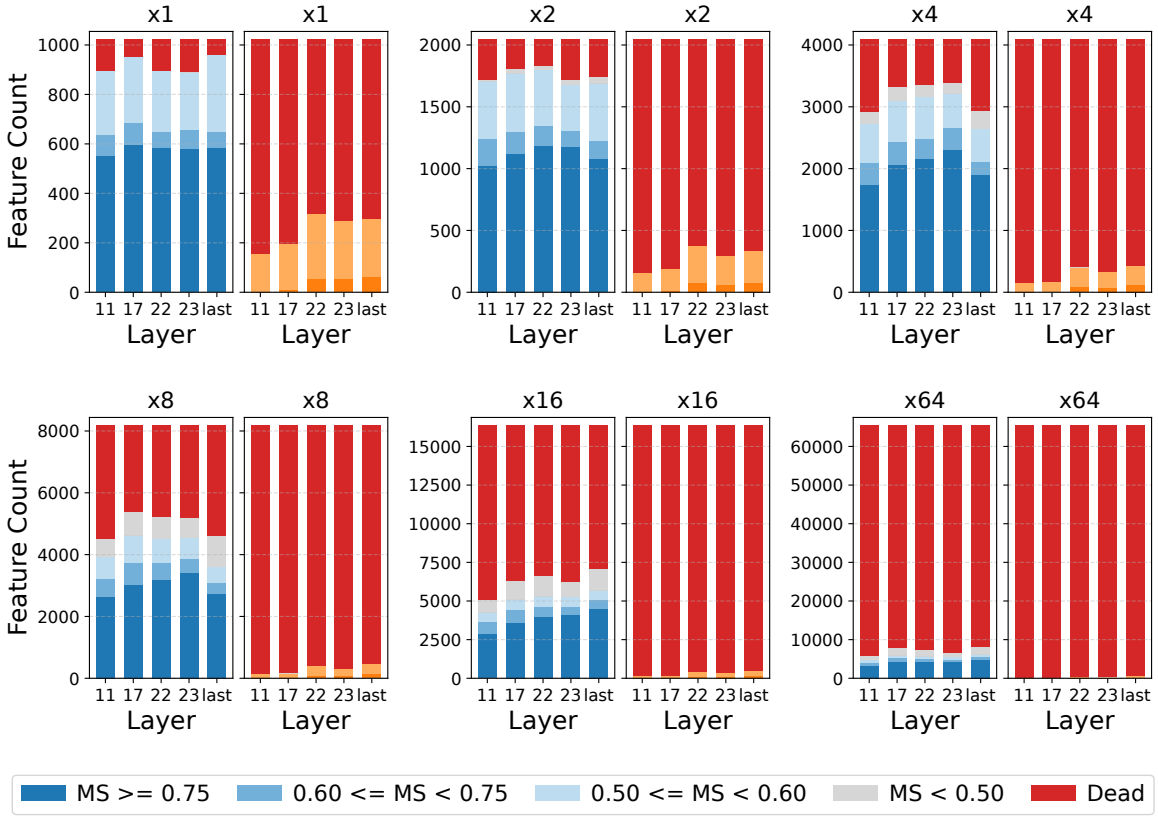

Figure 6. Comparison of the number of features on iNat2021 across five categories—dead features, MS < 0.5, 0.5–0.6, 0.6–0.75, and MS > 0.75—for FTC (blue) and TopK TC (orange) across expansion factors.

D. Number of Features in Three Monosemanticity Score Ranges

Additional results for layers 11, 22, and the last layer, grouped by monosemanticity score range, are presented below.

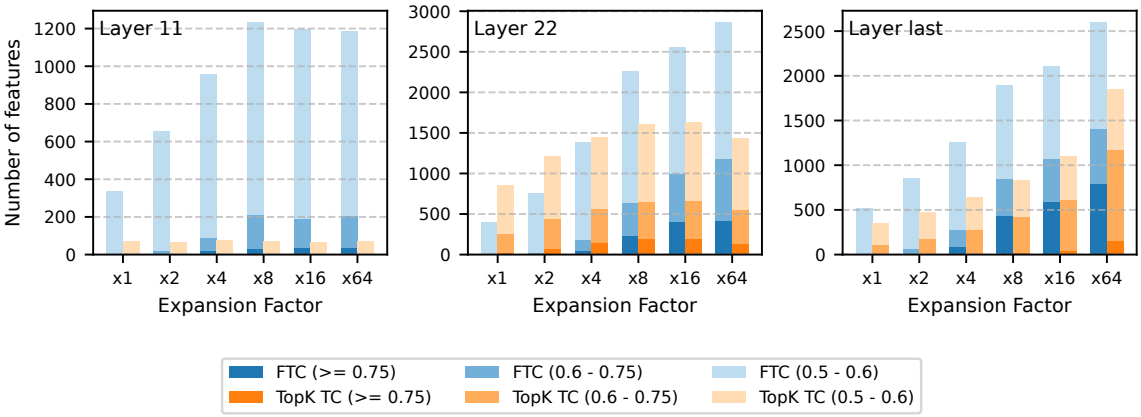

Figure 7. Number of features in three monosemanticity score ranges (0.5–0.6, 0.6–0.75, and > 0.75) for FTC (blue) and TopK TC (orange) across expansion factors, evaluated on ImageNet at remaining layers, 11, 22, and last.

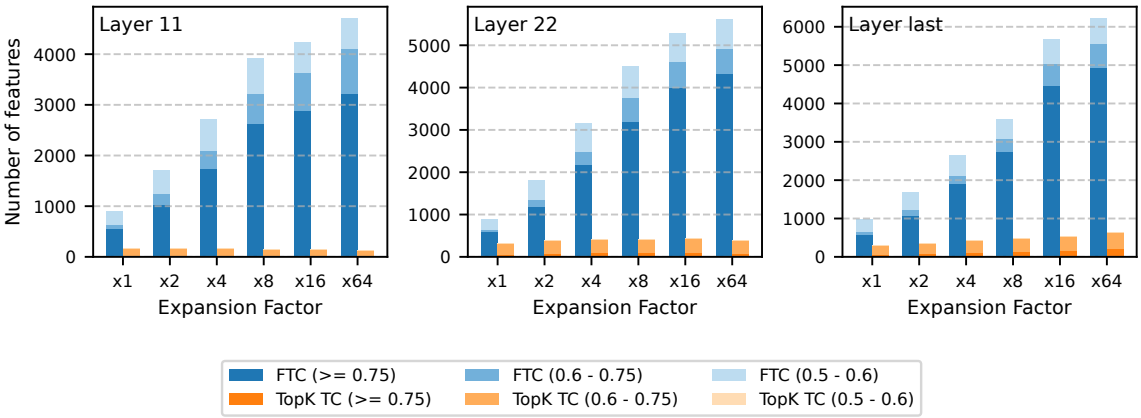

Figure 8. Number of features in three monosemanticity score ranges (0.5–0.6, 0.6–0.75, and > 0.75) for FTC (blue) and TopK TC (orange) across expansion factors, evaluated on iNat2021 at remaining layers, 11, 22, and last.

E. Feature Redundancy Analysis

During qualitative inspection of the visualizations, we observed that the top 16 activating images of some features were highly similar to those of other features (see Figure 9). Although the monosemanticity score is not computed solely from the top 16 images, such strong overlap suggests that these features may be closely correlated.

Figure 9. Visualization results of FTC on the ImageNet experiment at layer 11 with an expansion factor of 4. Three consecutive features with adjacent frequency values exhibit similar top-16 images.

To quantify this redundancy, we computed the Pearson correlation between feature activations and visualized a subset of the resulting correlation matrix as a heatmap. The heatmaps reveal that redundancy often appears between features associated with nearby frequencies.

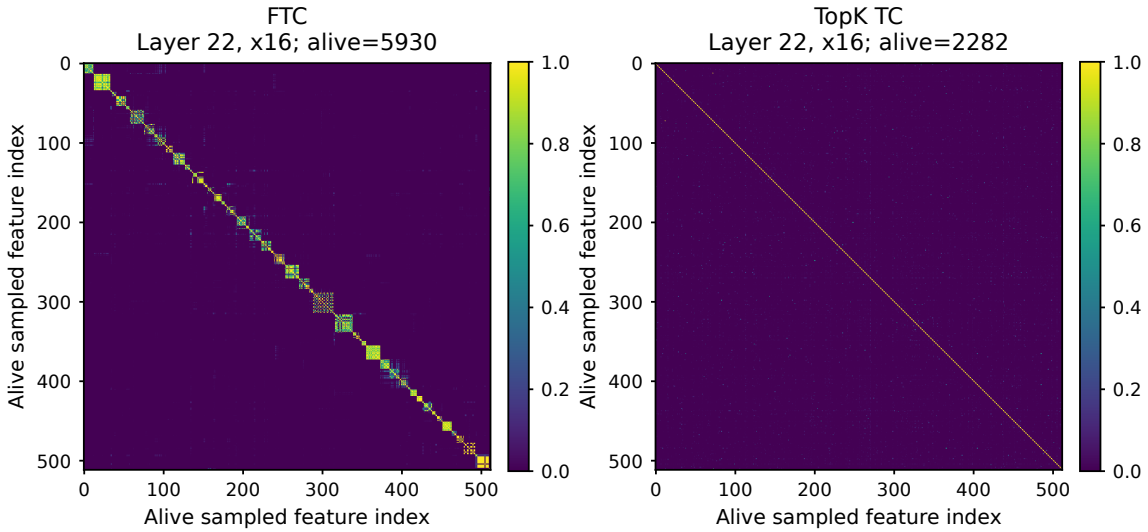

Figure 10. Pearson correlation heatmaps obtained from 512 sampled alive features at layer 22 with expansion factor 16. The left panel shows the result for FTC, and the right panel shows the result for TopK TC. Highly correlated blocks are observed in some neighboring-frequency regions, particularly in the yellow bands.

To further assess whether this redundancy substantially reduces the number of meaningful features, we performed an additional correlation-based analysis. On ImageNet, FTC and TopK TC become more comparable after accounting for feature redundancy, whereas on iNat2021, FTC still retains a substantially larger number of high-MS features. These results suggest that, although some redundancy exists among nearby Fourier features, FTC continues to provide a clear advantage in the number of effective high-quality features, especially on iNat2021.

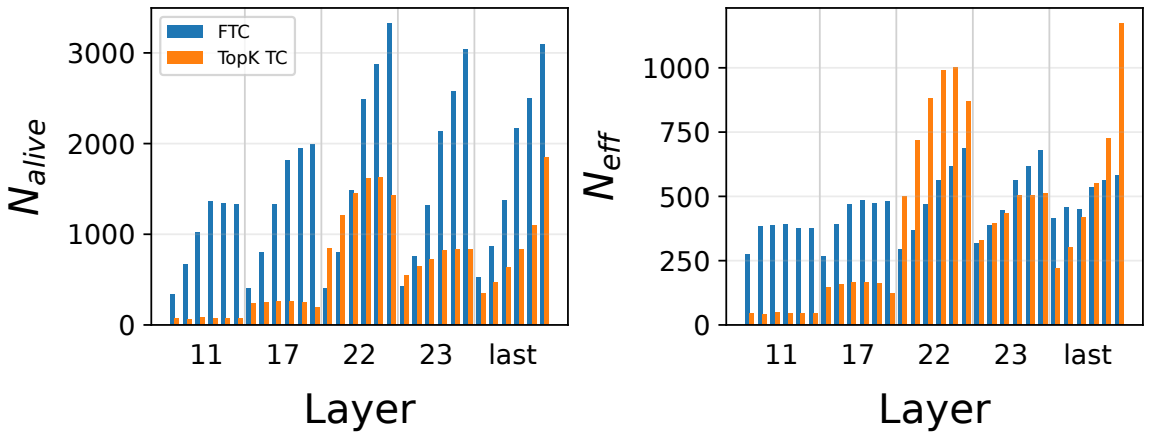

Figure 11. Results on ImageNet for the number of alive features (N_{alive}, left) and the effective number of features (N_{eff}, right). Although N_{eff} is reduced relative to N_{alive}, FTC remains higher than TopK TC, indicating that its advantage persists even after accounting for feature redundancy.

F. Union Coverage of High-MS Features on the Global t-SNE Embedding

We additionally conducted the union coverage analysis for Layers 17 and 23 at MS thresholds of 0.5 and 0.75. FTC consistently covered a broader range of images than TopK TC, even at the higher threshold, indicating that FTC's high-quality features are not concentrated in a narrow region but instead cover diverse images.

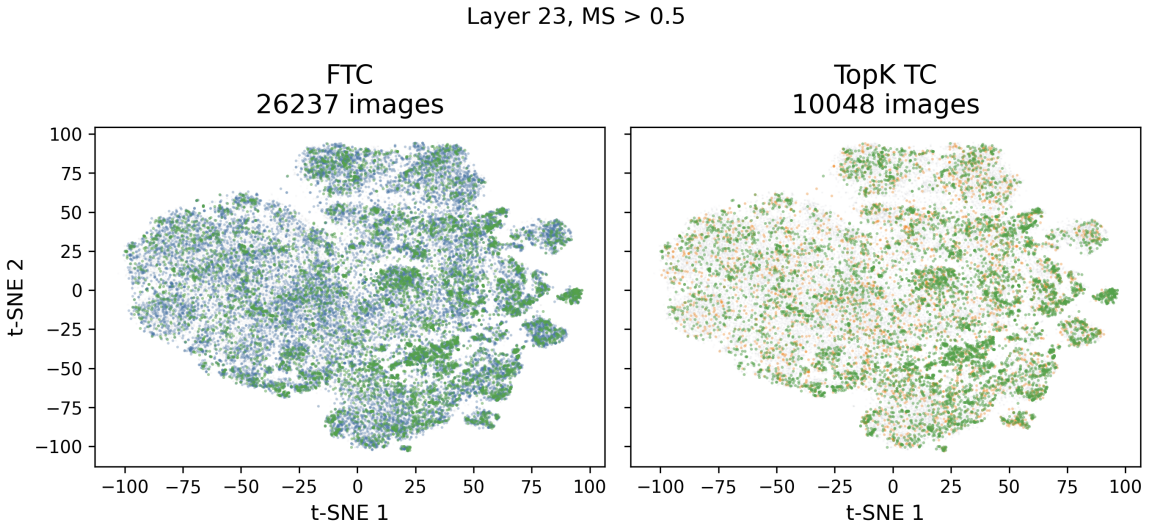

Figure 12. Union coverage of high-MS features on the global t-SNE embedding of the iNat validation set. For Layer 23 with expansion factor 16 and an MS threshold of 0.5, the panel shows the union of the top-30 activated images covered by FTC or TopK TC, with images covered by both models highlighted in green.

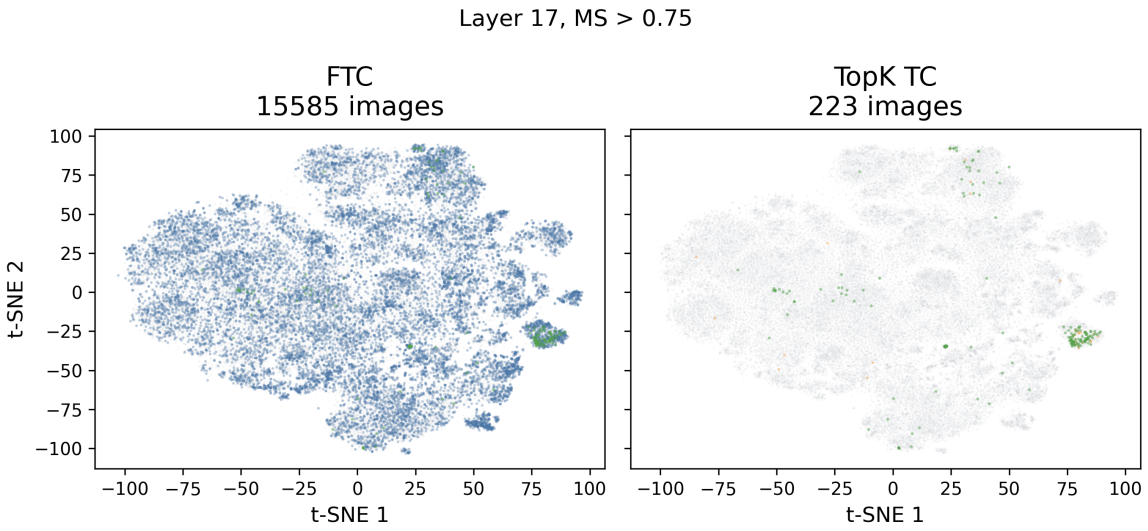

Figure 13. Union coverage of high-MS features on the global t-SNE embedding of the iNat validation set. For Layer 17 with expansion factor 16 and an MS threshold of 0.75, the panel shows the union of the top-30 activated images covered by FTC or TopK TC, with images covered by both models highlighted in green.

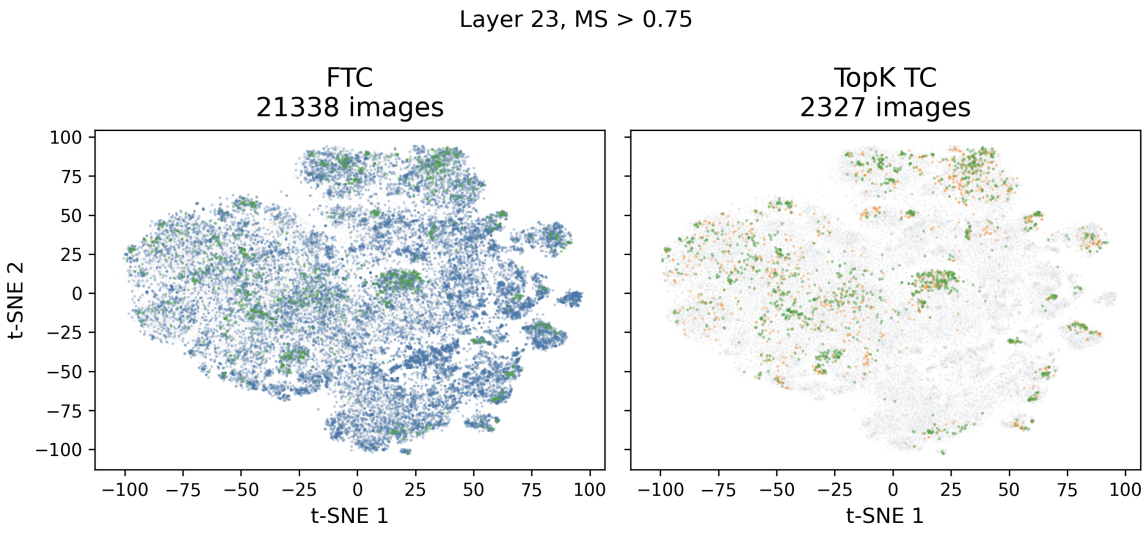

Figure 14. Union coverage of high-MS features on the global t-SNE embedding of the iNat validation set. For Layer 23 with expansion factor 16 and an MS threshold of 0.75, the panel shows the union of the top-30 activated images covered by FTC or TopK TC, with images covered by both models highlighted in green.