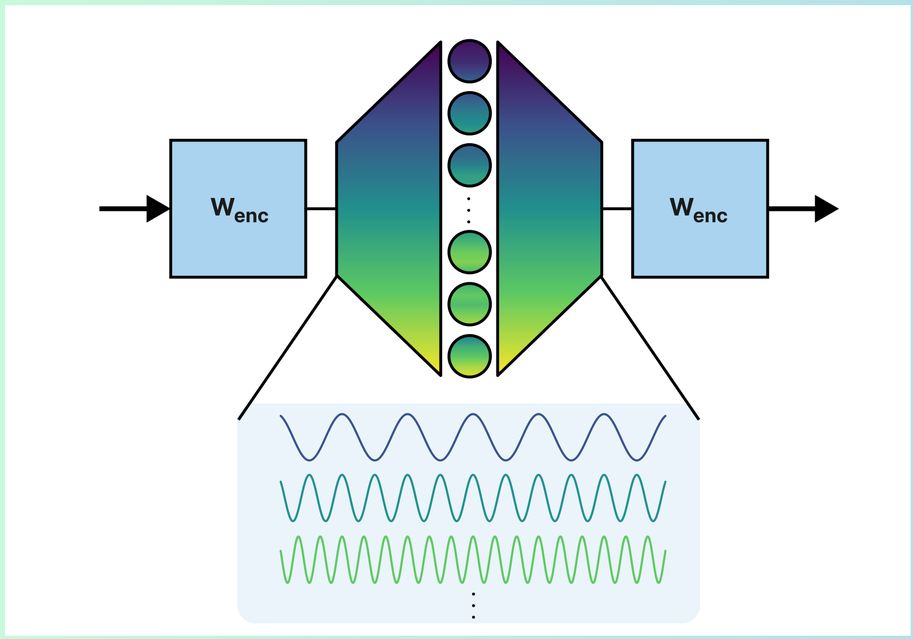


A Fourier basis enables transcoders to discover more monosemantic, active, and diverse visual features.

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

UNIVERSITY OF SEOUL

IDA.L Information Dynamics in Language Laboratory

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

TL;DR

- FTC uses a learnable **Fourier basis**, enabling **sparse and interpretable feature discovery** in CLIP MLP layers.
- FTC shows **fewer dead features** and **higher mono-semanticity** than TopK TC, with more usable features at comparable reconstruction quality.

Preliminaries

- Transcoders approximate an MLP sublayer as a sparse combination of interpretable features. Among TC variants, TopK TC is widely used, enforcing sparsity by keeping only the top-(k) latent activations. Formally:

$$z_{\mathrm{TC}}(x) = \mathrm{TopK}(\mathrm{ReLU}(W_{\mathrm{enc}}x + b_{\mathrm{enc}})),$$
$$\mathrm{TC}(x) = W_{\mathrm{dec}}z(x) + b_{\mathrm{dec}}.$$

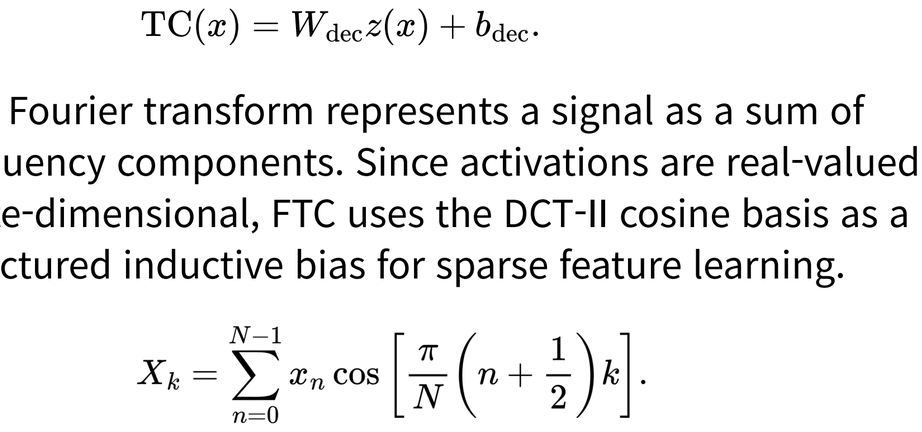
- The Fourier transform represents a signal as a sum of frequency components. Since activations are real-valued and finite-dimensional, FTC uses the DCT-II cosine basis as a structured inductive bias for sparse feature learning.

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

Fourier Transcoder

Fourier Transcoder (FTC)

- A TopK TC variant that uses a learnable Fourier basis in the encoder.
- Instead of selecting sparse features in a learned latent space, FTC selects them in a Fourier coefficient space.



Feature Examples

MS score = 0.719 MS score = 0.857 MS score = 0.896

Experiments

Monosemanticity (MS) Score

- How semantically similar the images that activate a feature are in an evaluation encoder's embedding space

- FTC yields **more high-MS features** than TopK TC, with a stronger gain on iNat2021.

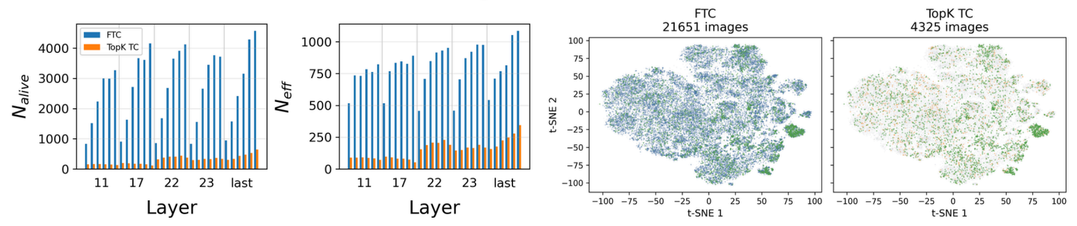
Reconstruction Quality

- How faithfully the Transcoder reconstructs the original MLP output, measured by R^2

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

Effective Feature Diversity & Coverage

- FTC discovers more effective and diverse high-MS features across the image space.



Limitations and Discussion

- FTC can produce partially redundant features, especially among nearby Fourier frequencies.
- Its reconstruction quality is competitive, but slightly lower than TopK TC in some settings.
- Does FTC's Fourier-basis decomposition also work for LLM activations beyond VLMs?

Questions & Discussion dylan16@uos.ac.kr