★ CVPR 2026 SPOTLIGHT ★ XAI4CV WORKSHOP

Can Cross-Layer Transcoders *replace* Vision Transformer activations? *An interpretable perspective on vision.*

Gerasimos Chatzoudis[1] · Konstantinos D. Polyzos[2] · Zhuowei Li[1]* · Difei Gu[1] · Gemma E. Moran[1] · Hao Wang[1] · Dimitris N. Metaxas[1]

[1] Rutgers University · [2] UC San Diego · * Work done outside of Amazon

TL;DR Cross-Layer Transcoders reconstruct and replace ViT MLP activations while preserving zero-shot accuracy, enabling faithful attribution of which layers build the final [CLS] and patch-token representations

Motivation

WHY INTERPRETABILITY

Vision Transformers achieve strong zero-shot performance, but their **internal representations** remain **difficult to interpret** [4]

MECHANISTIC INTERPRETABILITY

SAEs and **Transcoders** operate **within a single layer**, interpreting features **locally** but not explaining how information is transformed **across depth** [2,3]

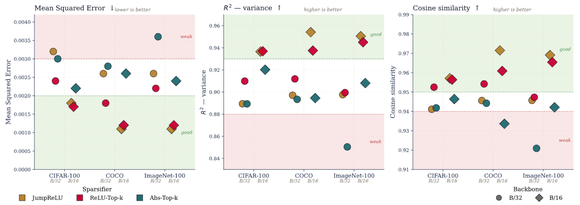
IDEA: CLTs FOR VISION

Leveraging the notion of CLTs from language to vision [1]: ViTs combine **spatial patch** tokens, **2D structure**, variable **granularity**, and a **global [CLS] representation**

Can CLTs faithfully replace ViT computations and reveal how visual representations are built across ViT depth?

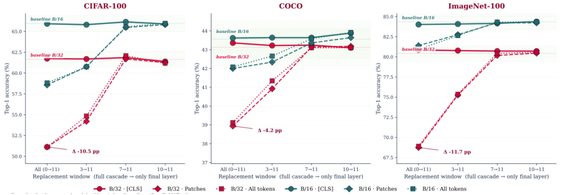
RQ 01 · FUNCTIONAL REPLACEMENT
Can CLTs functionally replace MLP blocks in ViTs as an alternative interpretable proxy?

Reconstruction of MLP Activations



Insight 01 · FUNCTIONAL REPLACEMENT based on Token Size
Patch granularity matters. ViT-B/16 reconstruction more faithful than ViT-B/32. Smaller patches distribute information across more tokens, yielding simpler per-token activations that are easier to approximate

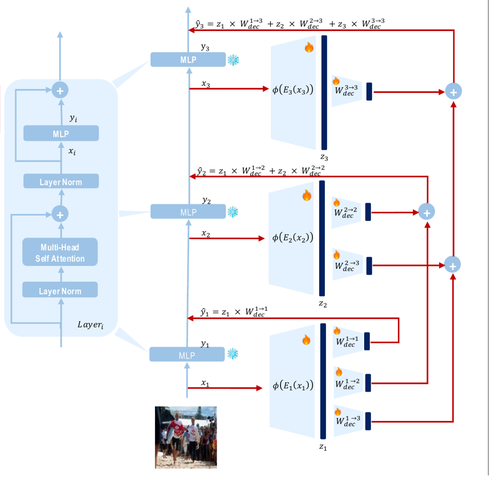
Cascaded Replacement



Insight 02 · FUNCTIONAL REPLACEMENT based on Token Type
CLTs can replace MLP blocks, especially in later layers or for [CLS] tokens across all layers, preserving and even improving in some cases zero-shot classification performance

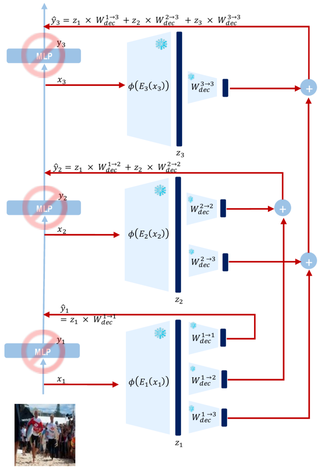
Training: *CLTs*

Use pre − MLP activations to predict target layer's post − MLP activation



Inference: *Replacement Model*

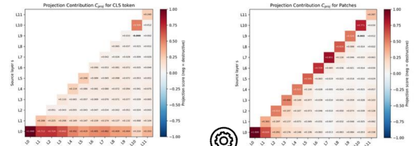
Replace MLP blocks with learned CLTs



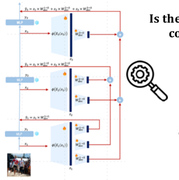
ViT Understanding through CLTs

Use CLTs as an interpretable proxy for ViTs

$$C_{i\to j}^{proj} = E_x\left[\left[\left(\frac{1}{T}\right)\sum_{t=1}^{T}\frac{\left\langle c_{i\to j}^{(t)}, \hat{y}_j^{(t)}(x)\right\rangle}{\left\|\hat{y}_j^{(t)}\right\|_2^2}\right]\right]$$



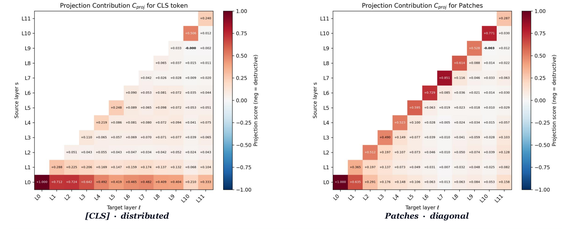
Is there a structural difference in cross − layer contributions between CLS and patches ?



Which layers are more significant in shaping the final representation?

RQ 02 · FAITHFUL ATTRIBUTION
Do cross-layer contribution scores yield faithful, process-level attribution?

CROSS-LAYER CONTRIBUTION SCORES



Insight 03 · CROSS-LAYER CONTRIBUTION SCORES based on Token Type
[CLS] integrates across depth; patches remain local. Patch tokens show diagonal attribution, while [CLS] draws credit from many preceding layers

FAITHULNESS ATTRIBUTION

			NECESSITY *remove top-1 attributed layer*		SUFFICIENCY *retain only top-4 layers*	
Dataset	**Tokens**	**Baseline**	**Drop Top-1**		**Keep Top-4**	
CIFAR-100	all	61.65	**59.68**	-1.97%	**59.96**	-1.69%
	[CLS]	61.65	**59.91**	-1.74%	60.72	-0.93%
COCO	all	43.12	**42.98**	-0.14%	43.02	-0.10%
	[CLS]	43.12	**42.78**	-0.34%	43.00	-0.12%
ImageNet-100	all	80.42	**74.94**	-5.48%	80.56	+0.14%
	[CLS]	80.42	**74.86**	-5.56%	80.48	+0.06%

Insight 04 · NECESSARY AND SUFFICIENT ATTRIBUTION LAYERS
The top-4 attributed layers recover accuracy, while removing the highest-scored layer causes substantial degradation

CROSS-LAYER FEATURE RETRIEVAL

CONCLUSION

CLTs make ViT computation **interpretable across depth**. They **preserve model behavior** while revealing that patch tokens are mostly layer-local, whereas [CLS] integrates information broadly across layers.

FUTURE WORK

From interpretation to intervention: extend CLTs to attention and larger VLMs, using sparse cross-layer features for circuit discovery and adversarial robustness.

REFERENCES

[1] Ameisen et al., Circuit tracing, 2025.
[2] Dunefsky et al., Transcoders find interpretable LLM feature circuits, 2024.
[3] Bricken et al., Towards monosemanticity, 2023.
[4] Radford et al., Learning transferable visual models from natural language supervision, 2021.

Corresponding Authors: Gerasimos Chatzoudis (gc745@cs.rutgers.edu) & Konstantinos D. Polyzos (kpolyzos@ucsd.edu)