Can Cross-Layer Transcoders Replace Vision Transformer Activations? An Interpretable Perspective on Vision

Supplementary Material

7. Visually Explainable Cross-Layer Contribution

The contribution scores $C_{s\to\ell}$ quantify *which* layers matter, but not *what* they represent. To obtain example-based explanations of cross-layer influence, we provide some evidence of a retrieval-based framework that operates directly on the CLT sparse codes. Recall that a target post-MLP representation $\hat{y}_L$ at the final layer can be decomposed as

$$\hat{y}_L = W^{\text{dec}}_{0\to L} z_0 + \cdots + W^{\text{dec}}_{L\to L} z_L$$

where each term corresponds to the contribution of a previous layer's sparse representation z_i, transformed by its decoder $W^{\text{dec}}_{i\to L}$. Instead of inspecting individual neurons z_{ik} in isolation, we treat the full sparse vector z_i as a feature and use it for layer-wise retrieval.

Concretely, we construct an external corpus from the training images of each dataset. For each image j and layer i, we pass the image through the CLT and extract its sparse codes $z_i^{(j)}$. We then aggregate token-level features (e.g., by averaging over spatial tokens) into a global sparse descriptor $z^{(j)}_{i,\text{agg}}$ and index these descriptors into a per-layer FAISS database D_i [9, 16]. At test time, for a given input image, we compute its sparse activations $\{z_i\}$ and use them to retrieve the top-K most similar training images from each layer's index:

$$\mathcal{N}_i(z_i) = \arg\operatorname*{topK}_j \text{sim}\left(z_i, z^{(j)}_{i,\text{agg}}\right), \quad \text{for } i \le \ell \qquad (12)$$

This mirrors the reconstruction process $\hat{y}^\ell = \sum_{i\le\ell} z_i W^{i\to\ell}_{\text{dec}}$ by surfacing images that share similar latent activations at each contributing layer. Unlike decoding, this retrieval provides interpretable, layer-specific visual evidence of what each layer's sparse features contribute to the final representation.

Figures 4 and 5 illustrate our retrieval-based framework for a single query image, showing the top-3 retrieved training samples per layer. For the `[CLS]` token, retrievals exhibit a depth-wise progression from low-level to high-level semantics: shallow layers primarily group images by generic color and layout statistics; mid-level layers emphasize similar configurations of people and objects; and the deepest layers retrieve highly class-consistent examples (e.g., images depicting the same activity, i.e., surfing), exhibiting robustness to viewpoint and background variation. This depth-wise semantic sharpening aligns with our cross-layer contribution analysis and with the strong functional performance of CLS-only substitution: as features become more class-specific with depth, CLT-based surrogates can both reconstruct and visually explain the representations driving the final decision.

Table 5. CLT surrogate faithfulness metrics.

	Pred. Distrib.			Top-k Agree		Embedding			Prompt Sens.	
Layers	ΔAcc	Flip↓	KL↓	Top-1	Top-5	Cos	CKA	Spear.	r	KL↓
0-11$^{\text{CLS}}$	+0.07	10.4%	.034	89.6	88.5	.984	.948	.929	.995	.036
7-11$^{\text{CLS}}$	+0.01	9.5%	.030	90.5	89.0	.985	.954	.937	.992	.031
10-11$^{\text{CLS}}$	−0.07	7.2%	.018	92.8	91.2	.991	.979	.973	.994	.019
11$^{\text{CLS}}$	−0.04	5.4%	.010	94.6	93.2	.996	.991	.991	.997	.010
0-11$^{\text{Patches}}$	−11.1	35.2%	.412	64.8	70.0	.957	.817	.759	.942	.383
7-11$^{\text{Patches}}$	+0.05	15.0%	.079	85.0	85.8	.990	.972	.957	.989	.073
10-11$^{\text{Patches}}$	−0.13	2.5%	.002	97.5	96.8	.999	.999	.998	.996	.002

8. Additional Metrics for CLTs' faithfulness

To provide further evidence of CLT's faithfulness to the original model, we evaluate distributional alignment (KL, flip rate), top-k agreement, embedding geometry (cos/CKA/Spearman), and prompt sensitivity across 18 templates for the ViT-B/32 model on CIFAR100. Table 5 shows that in the regimes the CLTs faithfully reconstruct the original model's representations (CLS across layers; late-layer replacement), the surrogate closely matches the teacher beyond accuracy (e.g., KL< 0.035, CKA> 0.94, prompt-trend $r > 0.98$). These results also validate the faithful CLT late-layer patch replacement (e.g., 10–11^P: KL$= 0.002$, flip$= 2.5\%$, CKA$= 0.999$). In contrast, early patch cascades degrade (e.g., 0–11^P: KL$= 0.412$, flip$= 35.2\%$), as identified in Table 3.

9. Training Details

Teacher models and datasets. All Cross-Layer Transcoders (CLTs) are trained on top of frozen CLIP image encoders with ViT-B/32 and ViT-B/16 backbones. For each backbone, we consider three datasets: CIFAR-100, COCO, and ImageNet-100. For every (dataset, backbone) pair we train three separate CLT variants, one for each sparsifier: JumpReLU, ReLU-Top-k, and Abs-Top-k (with $k = 128$). CLTs only access the internal activations of the teacher ViT and do not modify or finetune the underlying CLIP parameters.

Supervision and targets. Let $x_\ell \in \mathbb{R}^{T\times d}$ denote the post-attention (LN2), pre-MLP activations at layer ℓ, and let $y_\ell = \text{MLP}_\ell(x_\ell)$ be the corresponding post-MLP outputs. For each image, we run the frozen teacher ViT once and cache

Figure 4. Layerwise visual retrieval using CLT sparse codes of CLS for a test image. Each row shows the top-3 retrieved training samples across transformer layers, revealing the semantic evolution of representations.

Figure 5. Layerwise visual retrieval using CLT sparse codes of Patches for a test image. Each row shows the top-3 retrieved training samples across transformer layers, revealing the semantic evolution of representations.

(x_ℓ, y_ℓ) for all layers $\ell = 0, \ldots, 11$ and all tokens (both [CLS] and patch tokens). CLTs are trained to reconstruct y_ℓ from sparse features computed from $\{x_i\}_{i \le \ell}$, using teacher-forcing at training time; i.e., all CLT inputs come from the unmodified teacher trajectory.

Optimization hyperparameters. For all datasets, backbones, and sparsifiers we train CLTs with the AdamW optimizer, learning rate 2×10^{-4}, and an expansion factor of 16. Each CLT is trained for 10 epochs over the corresponding dataset, using all tokens (both [CLS] and patch tokens) in the loss. Hyperparameters are shared across datasets and backbones.

10. Reconstruction Accuracy across Layers

In Figures 6–10, we report the reconstruction quality of Cross-Layer Transcoders (CLTs) across all transformer layers on three datasets (CIFAR-100, COCO, and ImageNet-100) and two CLIP backbones (ViT-B/32 and ViT-B/16). For each configuration, we compare three sparsity variants, i.e., JUMPRELU, RELU-TOP-k, and ABS-TOP-k, using cosine similarity, mean squared error (log scale), and variance explained (R^2), averaged over all tokens in the test set.

11. Classification Accuracy under Cascaded CLT Replacement

We report top-1 classification accuracy (%) under cascaded CLT replacement across all layers ($s \to 11$) in Figures 6–23. For each dataset (CIFAR-100, COCO, ImageNet-100) and ViT backbone (ViT-B/32, ViT-B/16), we evaluate three sparsity mechanisms: JUMPRELU (JR), RELU-TOP-k (RTK), and ABS-TOP-k (ATK), under three token settings (CLS-only, patch-only, and all tokens). The replacement is performed progressively from early to late layers ($s = 0$ to $s = 11$), and results are compared to the frozen ViT baseline. We observe that across all datasets and backbones, CLS-token replacement achieves near-identical or slightly better accuracy compared to the original model. This shows that CLS tokens are robust to replacement. Regarding patch tokens, accuracy improves significantly as more layers are replaced, especially in the later layers.

12. Cross-Layer Contribution Scores

To better understand the internal attribution structure of Cross-Layer Transcoders (CLTs), we visualize in Figures 11–16 the contribution scores $C_{s \to \ell}$, which quantify the influence of each source layer s on the reconstruction of activations at target layer ℓ. These heatmaps reveal a clear depth-aware structure across all datasets and backbones, where contributions are strongest from temporally proximal layers.

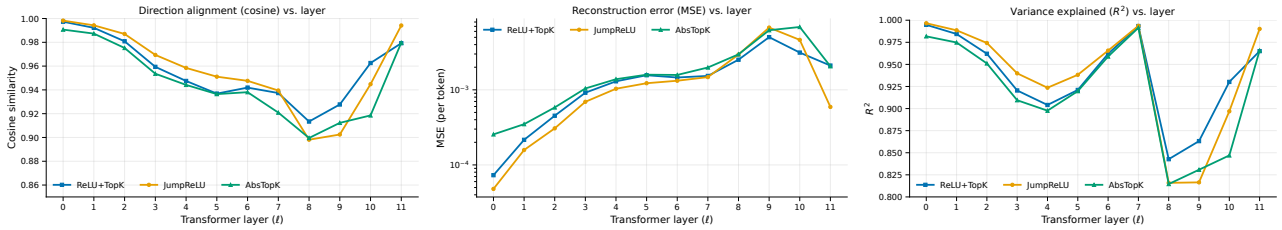


Figure 6. Reconstruction performance of CLTs across transformer layers on CIFAR-100 using ViT-B/16. We report cosine similarity (left), MSE per token in log scale (center), and R^2 (right) for JUMPRELU, RELU-TOP-k, and ABS-TOP-k sparsity variants (k=128), averaged across all tokens in the validation set.

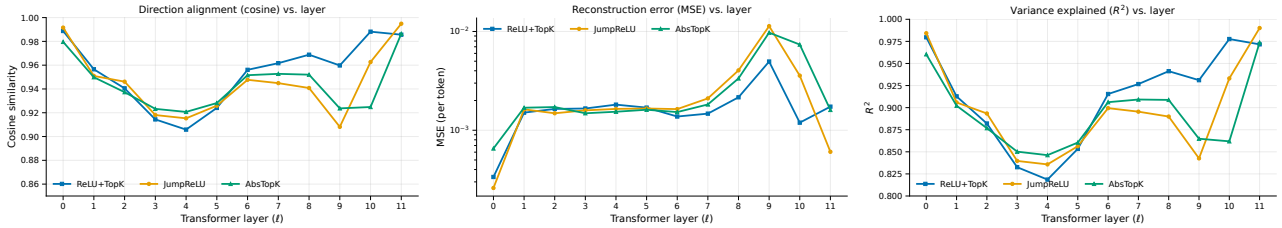


Figure 7. Reconstruction performance of CLTs across transformer layers on COCO using ViT-B/32. We show cosine similarity (left), MSE per token (log scale, center), and R^2 (right) for the three sparsity variants JUMPRELU, RELU-TOP-k, and ABS-TOP-k (k=128), averaged across all tokens in the validation set.

Table 6. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/32 with CLS tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	61.33	**61.74**	61.43
1→11	61.43	**61.76**	61.56
2→11	61.39	**61.72**	61.61
3→11	61.38	**61.69**	61.58
4→11	61.20	**61.76**	61.66
5→11	61.21	**61.59**	61.48
6→11	61.43	**61.58**	61.39
7→11	61.41	**61.86**	61.26
8→11	61.62	**61.90**	61.23
9→11	61.31	**61.85**	61.10
10→11	61.06	**61.39**	61.16
11→11	61.23	**61.31**	61.03

Notably, CLS tokens exhibit more distributed contributions spanning earlier layers, while patch tokens show sharper, more localized attribution.

Table 7. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/32 with Patches tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	49.63	**51.12**	48.68
1→11	49.92	51.61	**52.40**
2→11	51.95	52.86	**53.46**
3→11	53.65	54.20	**55.83**
4→11	56.59	57.02	**58.86**
5→11	58.89	59.52	**61.38**
6→11	60.39	61.16	**63.06**
7→11	61.51	61.69	**63.18**
8→11	61.83	62.04	**63.53**
9→11	61.82	61.73	**62.57**
10→11	61.49	61.25	**61.56**
11→11	**61.65**	**61.65**	**61.65**

Table 8. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/32 with All tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	49.40	**51.12**	48.41
1→11	49.90	51.91	**51.97**
2→11	51.84	**53.51**	53.16
3→11	54.10	54.84	**55.69**
4→11	56.62	57.32	**58.23**
5→11	59.21	59.96	**61.19**
6→11	60.54	61.60	**62.95**
7→11	61.54	62.08	**62.85**
8→11	61.69	62.32	**63.01**
9→11	61.63	61.83	**62.07**
10→11	61.01	**61.20**	61.11
11→11	61.23	**61.31**	61.03

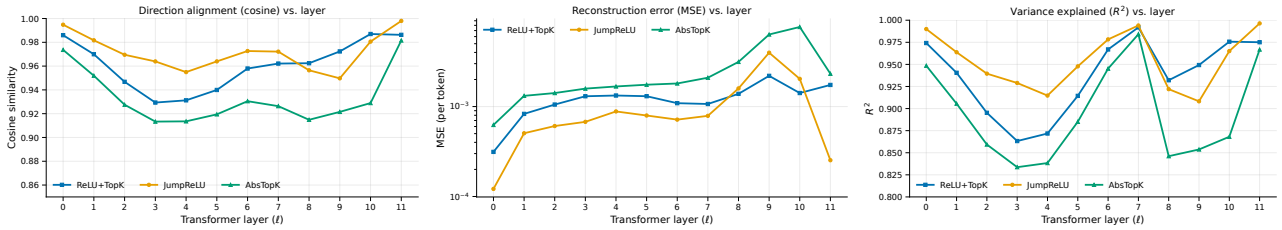


Figure 8. Reconstruction performance of CLTs across transformer layers on COCO using ViT-B/16. As in the main paper, we compare JUMPRELU, RELU-TOP-k, and ABS-TOP-k (k=128) with cosine similarity (left), MSE per token (log scale, center), and R^2 (right), averaged over the validation set.

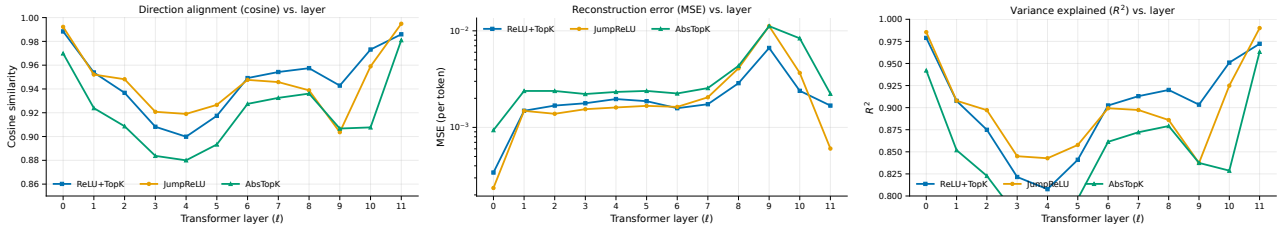


Figure 9. Reconstruction performance of CLTs across transformer layers on ImageNet-100 using ViT-B/32. We plot cosine similarity (left), MSE per token in log scale (center), and R^2 (right) for the three sparsity variants JUMPRELU, RELU-TOP-k, and ABS-TOP-k (k=128), averaged across all tokens in the validation set.

Table 9. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/16 with CLS tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	**66.04**	65.92	65.38
1→11	**66.05**	65.82	65.62
2→11	**66.02**	65.78	65.73
3→11	**66.10**	65.81	65.39
4→11	**66.05**	65.87	65.61
5→11	**66.15**	65.80	65.65
6→11	**66.16**	66.08	65.71
7→11	**66.15**	**66.15**	65.86
8→11	**66.12**	65.98	66.02
9→11	**66.12**	65.93	66.01
10→11	**66.06**	65.90	65.90
11→11	65.87	**66.00**	65.65

Table 10. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/16 with Patches tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	**62.40**	58.57	56.59
1→11	**62.58**	59.40	60.68
2→11	**63.15**	60.07	61.97
3→11	**63.54**	60.76	63.31
4→11	64.29	62.32	**64.85**
5→11	65.24	63.50	**66.11**
6→11	65.72	64.76	**66.85**
7→11	66.05	65.48	**67.24**
8→11	65.68	65.79	**66.79**
9→11	65.65	65.50	**66.06**
10→11	65.97	65.82	**66.13**
11→11	**65.97**	**65.97**	**65.97**

Table 11. Top-1 classification accuracy (%) on CIFAR-100 for ViT-B/16 with All tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 61.65 and ViT-B/16: 65.97.

Range	JR	RTK	ATK
0→11	**62.45**	58.82	56.72
1→11	**62.70**	59.23	60.30
2→11	**62.95**	59.74	61.54
3→11	**63.71**	60.75	62.87
4→11	64.14	62.16	**64.57**
5→11	65.20	63.60	**65.86**
6→11	66.08	64.73	**66.31**
7→11	66.22	65.56	**67.08**
8→11	65.72	65.84	**66.65**
9→11	65.77	65.51	**66.05**
10→11	65.91	**65.98**	65.89
11→11	65.87	**66.00**	65.65

Table 12. Top-1 classification accuracy (%) on COCO for ViT-B/32 with CLS tokens, across layer ranges $s \to 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	43.12	**43.36**	43.00
1→11	43.14	**43.32**	43.08
2→11	43.10	**43.26**	42.96
3→11	43.08	**43.22**	43.12
4→11	43.18	**43.22**	43.06
5→11	43.22	**43.30**	43.06
6→11	**43.30**	43.24	43.22
7→11	**43.26**	43.24	43.20
8→11	43.32	**43.40**	43.26
9→11	43.16	**43.34**	42.92
10→11	43.04	**43.10**	42.90
11→11	43.24	**43.26**	43.00

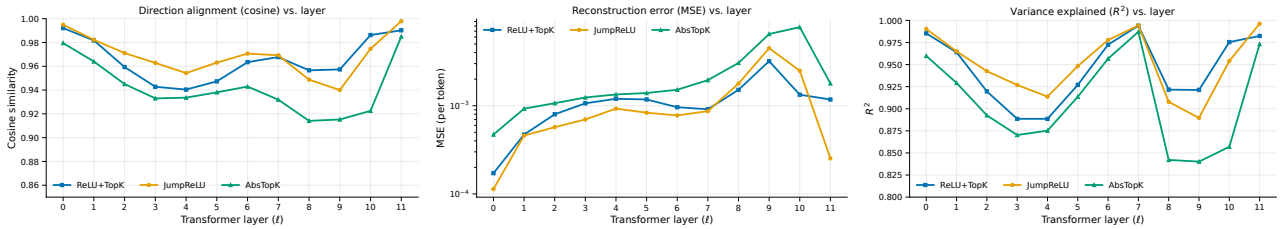
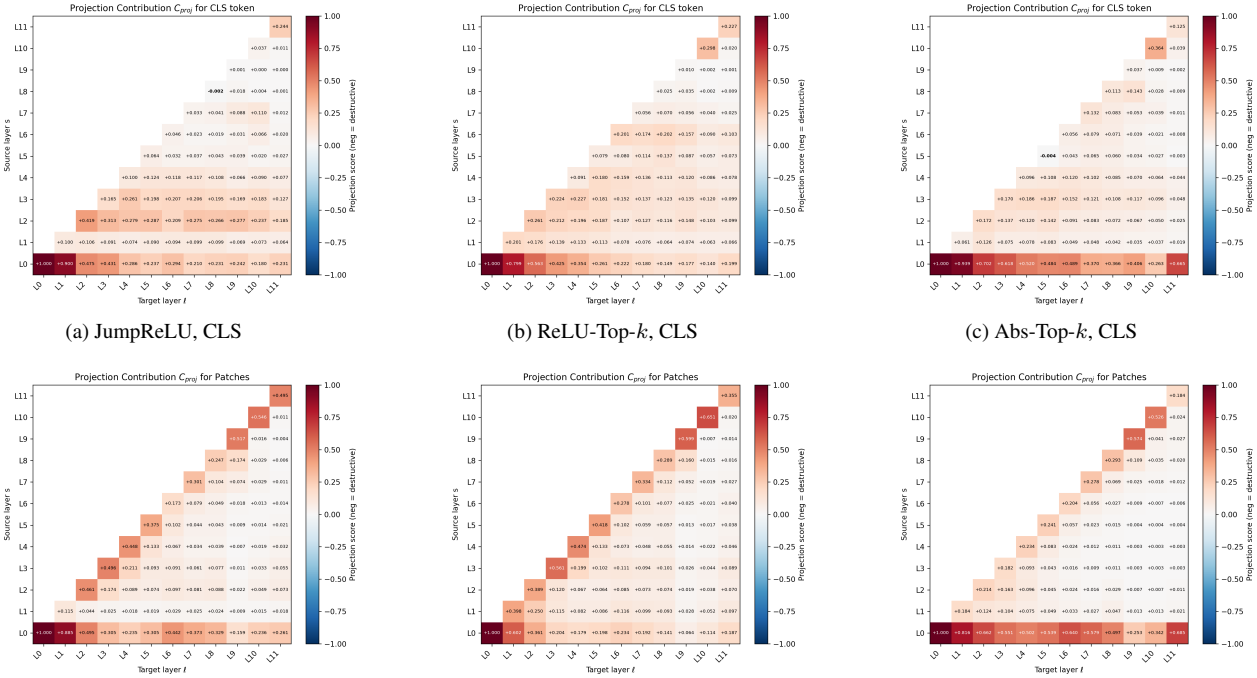


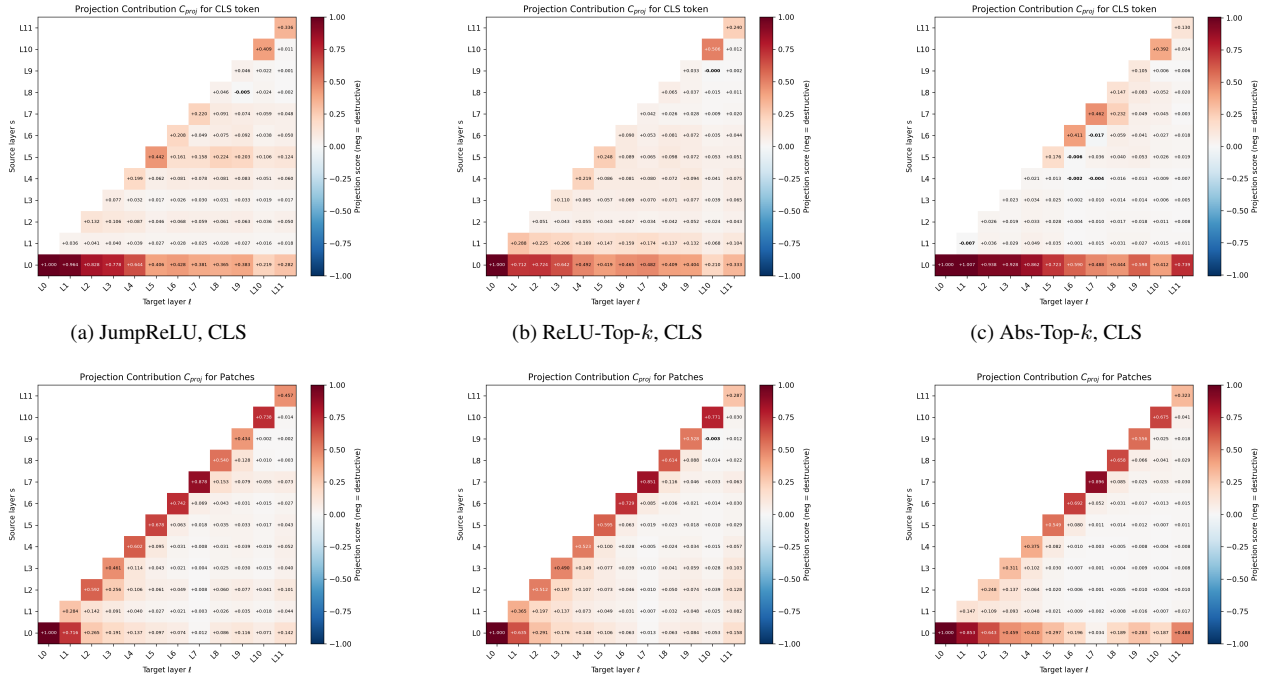
Figure 10. Reconstruction performance of CLTs across transformer layers on ImageNet-100 using ViT-B/16. Cosine similarity (left), MSE per token (log scale, center), and R^2 (right) are reported for JUMPRELU, RELU-TOP-k, and ABS-TOP-k sparsity (k=128), averaged across all tokens in the test set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

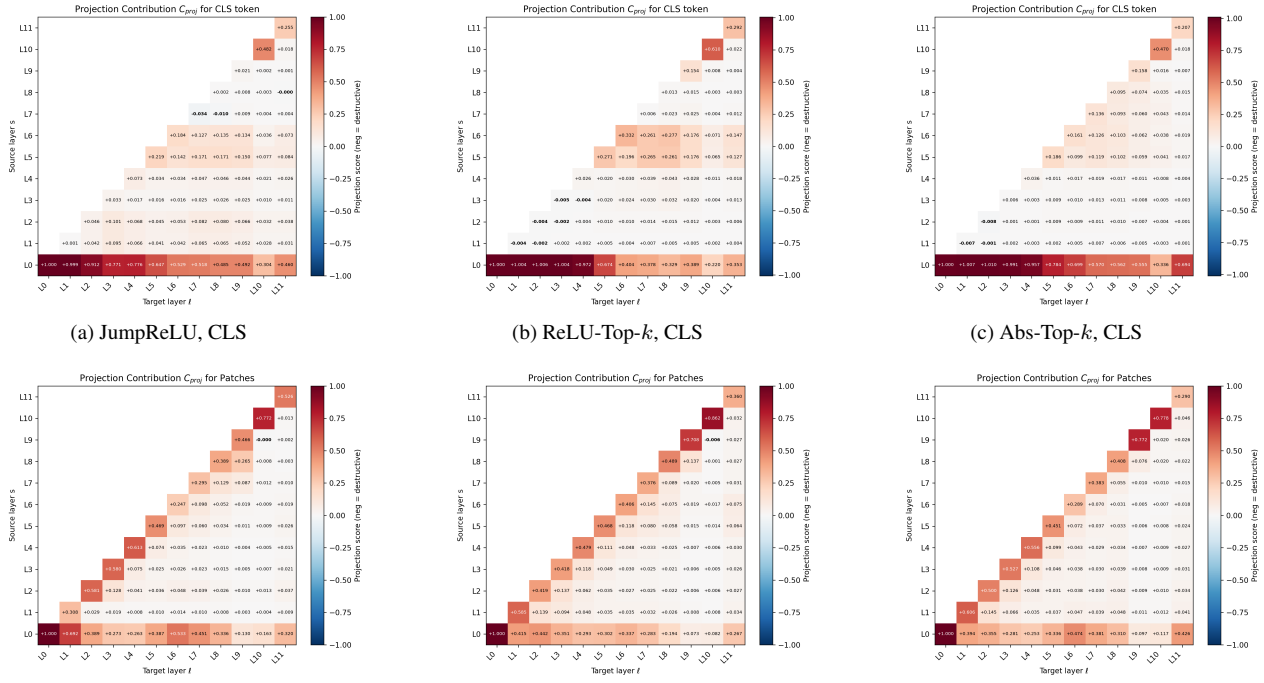
Figure 11. Cross-layer contribution scores $C_{s\to\ell}$ on CIFAR-100 with ViT-B/32. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

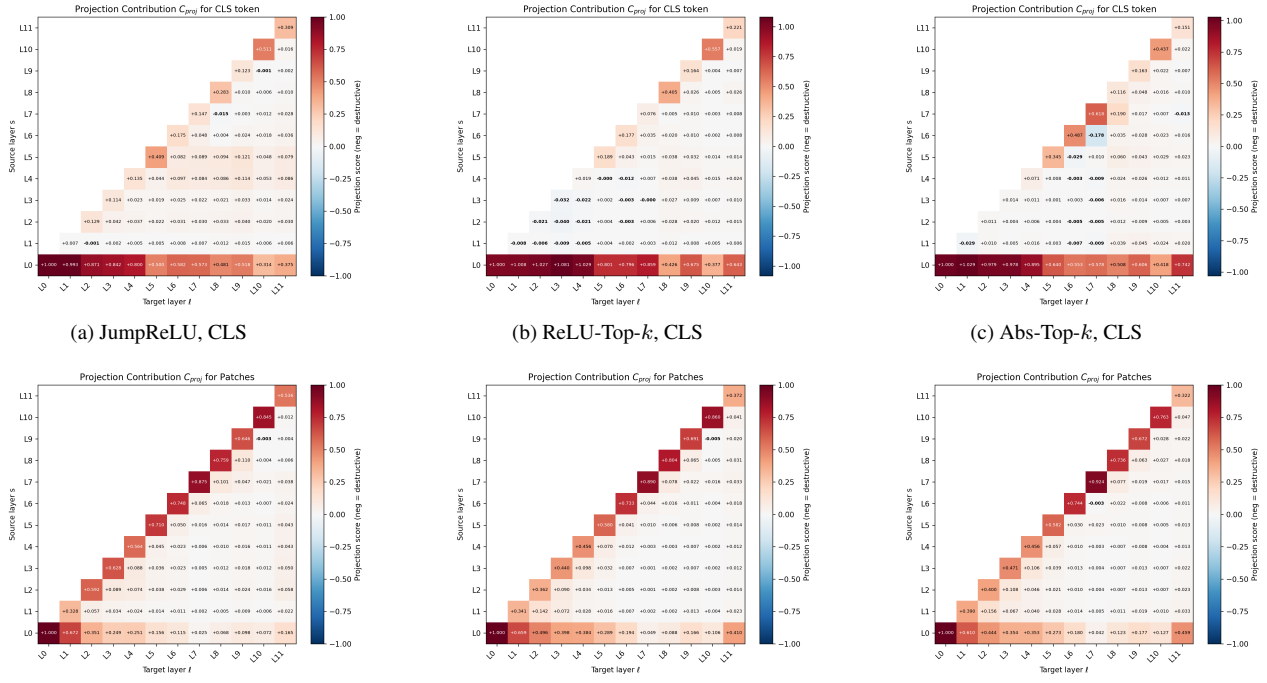
Figure 12. Cross-layer contribution scores $C_{s\to\ell}$ on CIFAR-100 with ViT-B/16. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

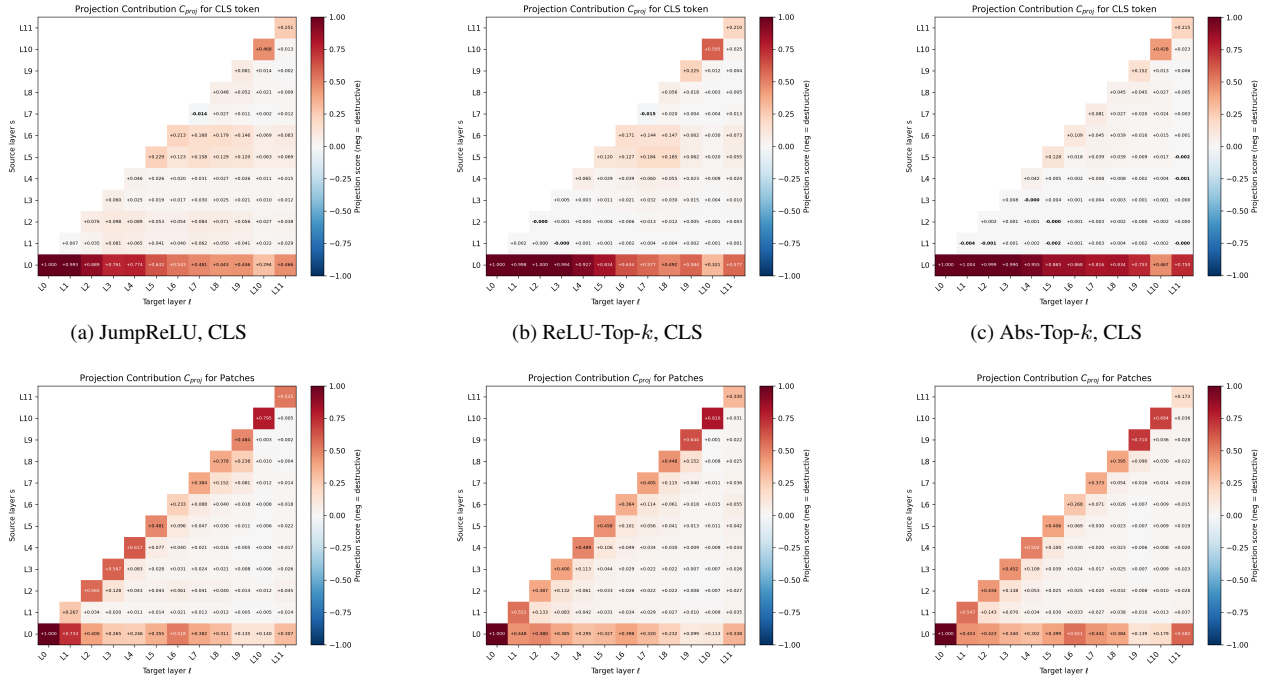
Figure 13. Cross-layer contribution scores $C_{s\to\ell}$ on COCO with ViT-B/32. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

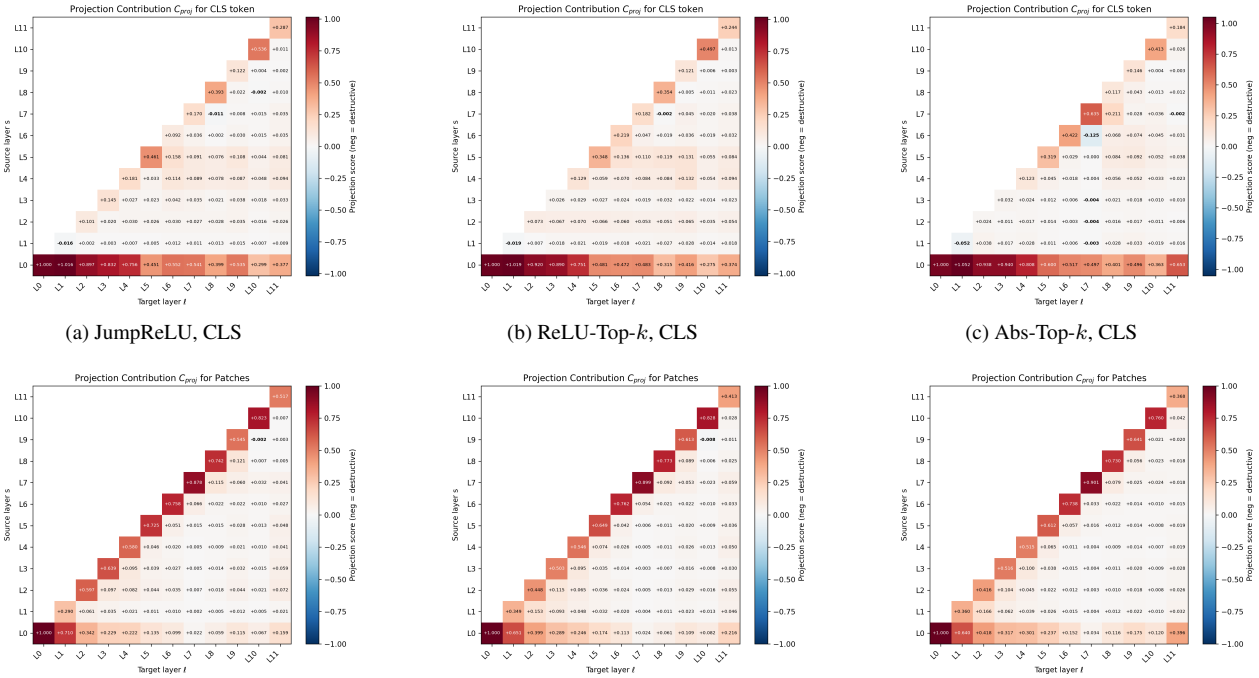
Figure 14. Cross-layer contribution scores $C_{s\to\ell}$ on COCO with ViT-B/16. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

Figure 15. Cross-layer contribution scores $C_{s\to\ell}$ on ImageNet-100 with ViT-B/32. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.



(a) JumpReLU, CLS (b) ReLU-Top-k, CLS (c) Abs-Top-k, CLS

(d) JumpReLU, patches (e) ReLU-Top-k, patches (f) Abs-Top-k, patches

Figure 16. Cross-layer contribution scores $C_{s\to\ell}$ on ImageNet-100 with ViT-B/16. Columns vary the sparsifier (JumpReLU, ReLU-Top-k, Abs-Top-k) and rows show CLS (top) and patch tokens (bottom). Each heatmap visualizes the proportional contribution of source layer s to the reconstructed activation at target layer ℓ, averaged over the validation set.

Table 13. Top-1 classification accuracy (%) on COCO for ViT-B/32 with Patches tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	39.04	38.94	**39.68**
1→11	39.38	39.44	**40.02**
2→11	39.54	40.24	**40.42**
3→11	40.58	40.92	**41.48**
4→11	41.38	41.88	**42.10**
5→11	42.06	**42.28**	**42.28**
6→11	42.48	42.58	**42.60**
7→11	42.98	**43.14**	42.80
8→11	**43.14**	43.02	42.92
9→11	**43.14**	42.92	42.92
10→11	43.14	**43.18**	43.06
11→11	**43.12**	**43.12**	**43.12**

Table 14. Top-1 classification accuracy (%) on COCO for ViT-B/32 with All tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	38.60	39.12	**40.06**
1→11	39.02	39.62	**40.54**
2→11	39.98	40.10	**40.44**
3→11	41.00	41.34	**41.46**
4→11	41.62	41.68	**41.92**
5→11	41.88	**42.28**	42.26
6→11	**42.88**	42.60	42.82
7→11	**43.14**	43.10	42.76
8→11	**43.26**	**43.26**	42.86
9→11	**43.28**	43.20	42.74
10→11	42.78	**43.10**	42.92
11→11	43.24	**43.26**	43.00

Table 15. Top-1 classification accuracy (%) on COCO for ViT-B/16 with CLS tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	**43.62**	**43.62**	42.76
1→11	43.56	**43.62**	42.40
2→11	43.60	**43.64**	42.34
3→11	**43.68**	43.64	42.64
4→11	43.66	**43.72**	42.64
5→11	43.58	**43.80**	42.98
6→11	43.56	**43.66**	43.50
7→11	43.52	**43.64**	43.34
8→11	**43.72**	43.66	43.68
9→11	43.82	**43.84**	43.24
10→11	43.52	**43.88**	43.38
11→11	43.50	**43.62**	43.28

Table 16. Top-1 classification accuracy (%) on COCO for ViT-B/16 with Patches tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	**42.72**	42.00	35.46
1→11	**42.84**	41.98	37.28
2→11	**42.68**	42.20	38.36
3→11	**43.00**	42.34	39.46
4→11	**43.00**	42.96	40.86
5→11	42.98	**43.00**	42.58
6→11	**43.38**	43.16	42.98
7→11	43.26	**43.36**	43.30
8→11	43.44	43.56	**43.86**
9→11	43.62	43.70	**43.76**
10→11	43.62	43.64	**43.72**
11→11	**43.50**	**43.50**	**43.50**

Table 17. Top-1 classification accuracy (%) on COCO for ViT-B/16 with All tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 43.12 and ViT-B/16: 43.56.

Range	JR	RTK	ATK
0→11	**43.00**	42.08	34.16
1→11	**42.92**	42.08	36.06
2→11	**42.74**	42.20	36.82
3→11	**42.78**	42.66	38.18
4→11	42.84	**43.10**	39.80
5→11	43.04	**43.12**	42.22
6→11	43.44	**43.56**	43.04
7→11	**43.62**	43.58	43.48
8→11	**43.72**	43.70	43.66
9→11	43.66	**43.96**	43.56
10→11	43.42	**43.92**	43.60
11→11	43.50	**43.62**	43.28

Table 18. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/32 with CLS tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**80.92**	80.86	80.26
1→11	**80.84**	80.82	80.32
2→11	80.82	**80.86**	80.24
3→11	**80.86**	80.78	80.16
4→11	80.80	**80.86**	80.24
5→11	80.74	**80.86**	80.44
6→11	80.68	**80.84**	79.92
7→11	80.64	**80.72**	80.16
8→11	80.66	**80.68**	80.06
9→11	80.46	**80.60**	80.10
10→11	80.56	**80.72**	80.18
11→11	**80.54**	**80.54**	80.36

Table 19. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/32 with Patches tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**71.96**	68.74	60.26
1→11	**72.34**	70.10	64.80
2→11	**73.10**	72.24	67.54
3→11	**75.58**	75.26	72.46
4→11	**77.86**	77.18	75.88
5→11	78.92	**79.10**	77.54
6→11	**79.80**	79.78	79.56
7→11	80.34	80.18	**80.64**
8→11	80.60	80.38	**80.86**
9→11	80.62	80.50	**80.76**
10→11	**80.52**	80.46	80.46
11→11	**80.42**	**80.42**	**80.42**

Table 20. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/32 with All tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**71.60**	68.90	60.26
1→11	**71.92**	70.10	63.82
2→11	**72.88**	72.32	66.72
3→11	75.12	**75.36**	71.44
4→11	**77.56**	77.10	74.68
5→11	78.74	**78.84**	77.30
6→11	79.50	**79.86**	78.82
7→11	80.18	**80.44**	79.38
8→11	**80.52**	80.24	80.32
9→11	**80.72**	80.48	80.46
10→11	80.38	**80.58**	80.18
11→11	**80.54**	**80.54**	80.36

Table 21. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/16 with CLS tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**84.54**	84.02	83.28
1→11	**84.50**	84.04	83.52
2→11	**84.44**	83.98	83.50
3→11	**84.56**	84.08	83.74
4→11	**84.44**	84.04	83.82
5→11	**84.42**	84.10	84.20
6→11	**84.52**	84.04	84.24
7→11	**84.66**	84.16	84.40
8→11	**84.60**	84.30	84.46
9→11	**84.62**	84.30	84.42
10→11	**84.56**	84.38	84.34
11→11	**84.46**	84.36	84.36

Table 22. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/16 with Patches tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**83.04**	81.40	73.06
1→11	**83.00**	81.44	76.74
2→11	**83.18**	81.88	77.54
3→11	**83.16**	82.76	79.36
4→11	**83.56**	83.34	80.98
5→11	83.78	**83.88**	82.50
6→11	**84.20**	84.12	83.26
7→11	83.94	**84.24**	83.84
8→11	84.02	84.24	**84.26**
9→11	84.20	84.22	**84.72**
10→11	84.20	84.22	**84.58**
11→11	**84.34**	**84.34**	**84.34**

Table 23. Top-1 classification accuracy (%) on ImageNet-100 for ViT-B/16 with All tokens, across layer ranges $s \rightarrow 11$. JR = JumpReLU, RTK = ReLU-Top-k, ATK = Abs-Top-k. Baseline top-1 accuracies (in %) are ViT-B/32: 80.42 and ViT-B/16: 84.34.

Range	JR	RTK	ATK
0→11	**83.12**	80.94	70.78
1→11	**83.08**	81.58	74.70
2→11	**83.28**	81.90	76.30
3→11	**83.00**	82.64	78.08
4→11	**83.62**	83.08	80.12
5→11	**84.16**	83.86	81.76
6→11	**84.46**	83.92	82.56
7→11	84.28	**84.32**	83.34
8→11	**84.38**	84.10	84.08
9→11	84.20	84.22	**84.74**
10→11	84.40	84.36	**84.74**
11→11	**84.46**	84.36	84.36