NEURAL INFORMATION PROCESSING SYSTEMS | NEURIPS 2026 WORKSHOP · ON-DEVICE INTELLIGENCE: FOUNDATION MODELS UNDER REAL-WORLD CONSTRAINTS

Which Transformer Blocks Should You Keep?

An Exhaustive Map of Layer-Subset Selection in a Pretrained ViT

Try me! →

Pick your own blocks
skiplayer.jung.vision

Steffen Jung · Margret Keuper
University of Mannheim · Max Planck Institute for Informatics, Saarland Informatics Campus
steffen.jung@uni-mannheim.de

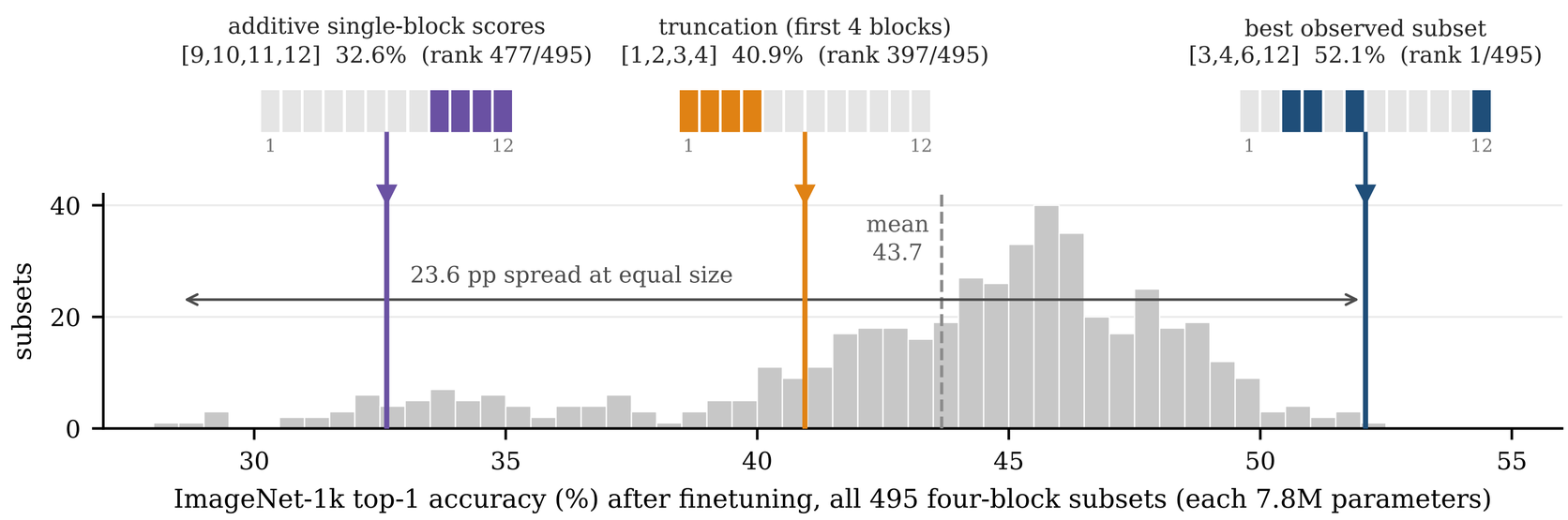
UNIVERSITÄT MANNHEIM

MAX PLANCK INSTITUTE FOR INFORMATICS

Same parameter count — 23.6 points apart

Which four would you keep?
Pick them on the project page ↗

All 495 four-block subsets have exactly 7.8M parameters. After identical finetuning their ImageNet-1k accuracy spans **23.6 pp**. Keeping the first four blocks (**truncation**) or the blocks with the smallest single-deletion cost (**additive scores**) both land below the mean.



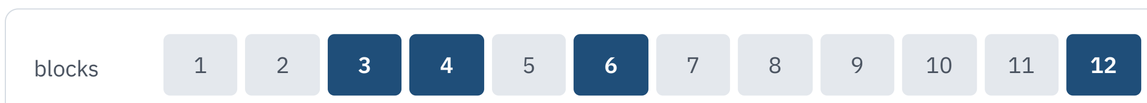
TAKE-HOME

1. **Don't truncate by default.** Keeping the first k blocks is below the subset mean at every pruned depth; every other subset beats it at k = 10, 11.
2. **Choose blocks for the target budget, jointly.** Block value flips with budget, and deletions interact: single-block scores anti-correlate with accuracy at k = 4.
3. **A small search goes far.** Local search lands within 0.18 pp of the best subset after 50 evaluations; truncation is 8.5 pp short.

Benchmark: the complete measured landscape is public, so block-selection methods can be checked against the best subset at every budget.

1 Setup: measure every subset

Dropping whole blocks keeps execution dense — no sparse kernels — and cuts parameters and FLOPs together. The question is **which** blocks to keep.

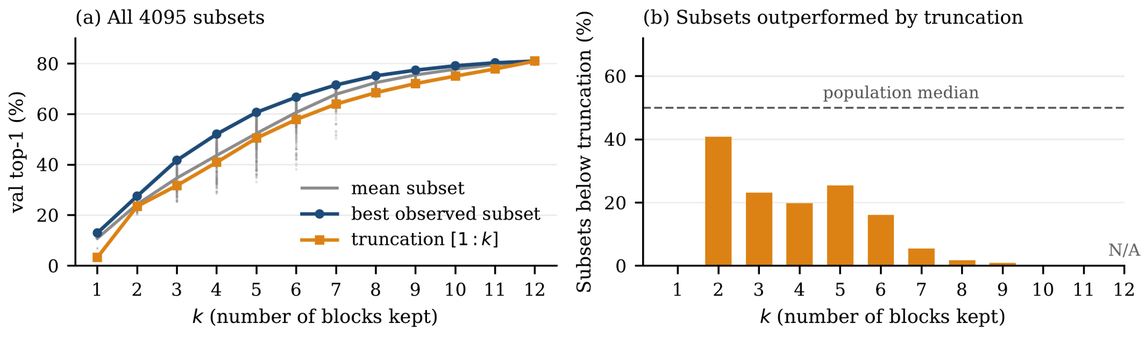


A subset is a mask $s \in \{0,1\}^{12}$; kept blocks run once, in their original order — here [3, 4, 6, 12], k = 4.

- **Model:** DINOv3 ViT-S/16, 12 blocks, 21.98M parameters, 80.96% top-1 on ImageNet-1k.
- **Space:** all $2^{12} - 1$ = **4,095** nonempty subsets, each finetuned end-to-end for 3 epochs with a fresh head.
- **Size is fixed by k:** 2.458M + 1.775M·(k–1) parameters, whichever blocks are kept.
- **Baseline: truncation** keeps the first k blocks, [1, …, k].

2 Truncation is a poor default

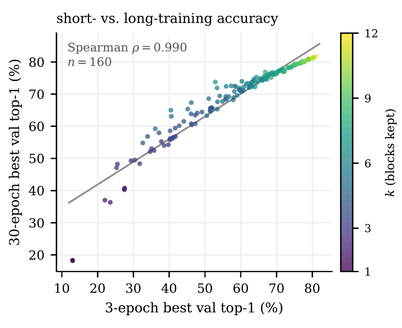
Truncation is **below the subset mean at every pruned depth**, so a uniformly random subset has higher expected accuracy. At k = 7–9 at least 94% of subsets beat it.



k (blocks kept)	1	2	3	4	5	6	7	8	9	10	11
best observed	13.0	27.6	41.8	52.1	60.7	66.7	71.5	75.2	77.4	79.1	80.3
subset mean	10.9	24.3	34.7	43.7	52.4	60.7	67.9	72.5	75.5	77.8	79.6
truncation	3.3	23.5	31.8	41.0	50.5	57.9	64.0	68.5	72.1	75.1	77.9
best – trunc. (pp)	+9.7	+4.0	+10.0	+11.1	+10.2	+8.8	+7.6	+6.7	+5.3	+4.1	+2.4
trunc. percentile	0	41	23	20	25	16	5	2	1	0	0

ImageNet-1k val top-1 (%) over all (12 choose k) subsets of each size.

The best **4-block** model (52.1%, 7.8M params) beats the worst **7-block** model (50.4%, 13.1M) with 41% fewer parameters — **which** blocks can matter more than **how many**.

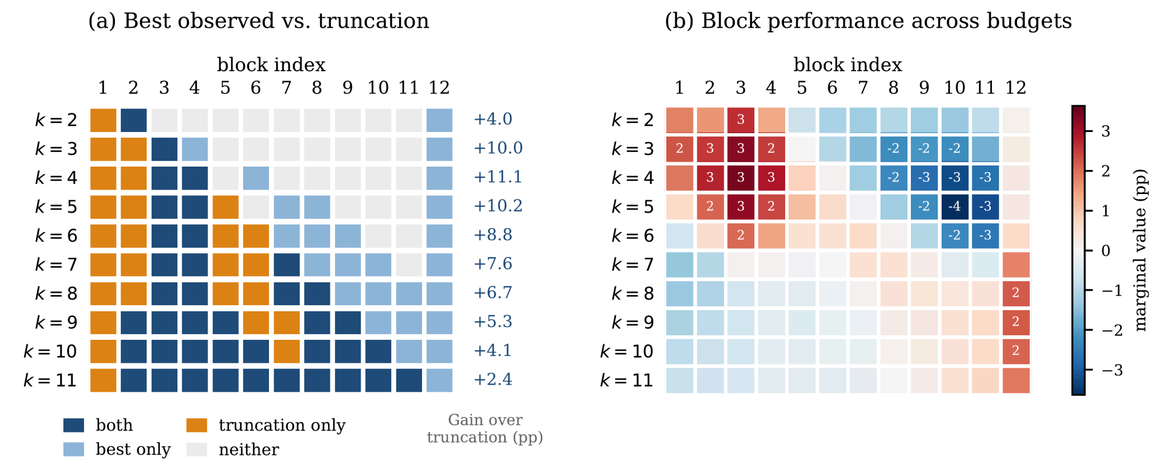


Is 3-epoch finetuning a fair proxy?

For 160 subsets also trained for 30 epochs, accuracies agree with Spearman ρ = 0.99.

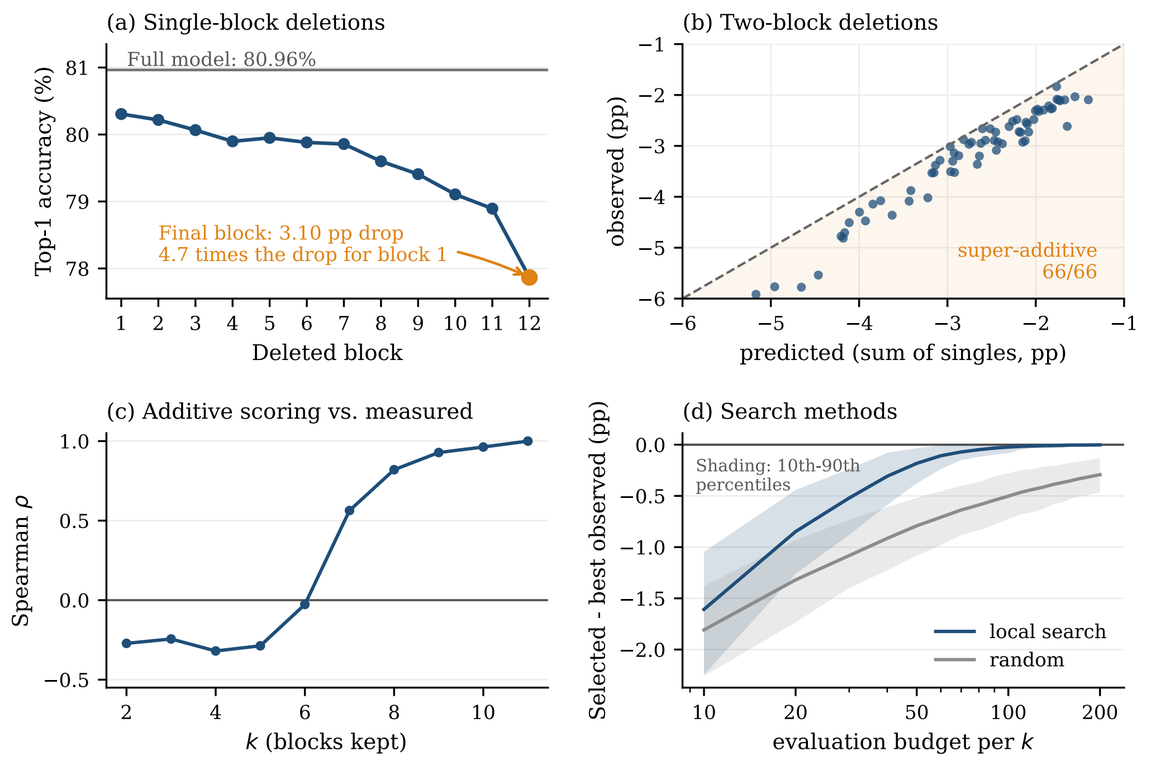
Finetuning is needed, though: with a frozen backbone, removing a single block already costs ≥ 7.6 pp versus finetuning the same subset.

3 Good subsets mix early and late blocks



- **Recurring choice:** for every k = 3…11 the best subset keeps blocks 3, 4 and 12 and drops block 1. At k = 4 it is [3, 4, 6, 12] instead of [1, 2, 3, 4]: **+11.1 pp**.
- **Not nested:** at 4 of 11 steps $k \to k$+1 a block leaves the best subset (e.g. block 6 is swapped for 7 and 8). Add-only or remove-only procedures cannot recover every best subset.
- **Importance depends on the budget** (b): a block's marginal value $m_i(k)$ (mean accuracy of k-subsets with vs. without it) flips sign, e.g. $m_3(4)$ = +3.45 but $m_3(8)$ = −0.61 pp. Only block 12 helps throughout, so **a full-model ranking need not transfer to tight budgets**.

4 Single-block scores mislead; search works



- **(a)** Single-block deletion cost grows with depth: 0.66 pp (block 1) to 3.10 pp (block 12).
- **(b)** For **all 66 pairs**, deleting both hurts more than the two single deletions combined.
- **(c)** Additive scores rank subsets well at k = 10 (ρ = 0.96) but inversely at k = 4 (ρ = −0.32).
- **(d)** Swap-based local search is within **0.18 pp** of the best subset after 50 evaluations (random sampling: 0.79 pp).

Model: DINOv3 (Siméoni et al., 2025) · Data: ImageNet-1k (Russakovsky et al., 2015) · Related block-pruning heuristics: ShortGPT (Men et al., 2024), SLEB (Song et al., 2024), Gromov et al. (2025)

Supported by BMFTR project TrackOpt (01IS24074A-D) and DFG Research Unit 5336 – Learning to Sense