AERO × VISION

The training-loop advantage carries across five vision tasks — medical and natural images alike.

First measured on large-language-model pretraining, AERO now beats a **tuned** AdamW baseline on **five public vision benchmarks** — spanning medical segmentation, medical classification, and the canonical CIFAR natural-image tasks — in two currencies at once: **higher accuracy** and **less compute** to reach it. In every comparison, only the optimizer changes.

5 / 5
Vision tasks won vs tuned AdamW
every properly-tuned head-to-head we've run

+1.0–3.3 pp
Higher accuracy / Dice at peak
across all five benchmarks

2–4×
Fewer steps to reach AdamW's peak
2.1× to 4.2× compute-to-target

BENCHMARK	TASK · MODEL	ACCURACY / DICE GAIN	STEPS TO ADAMW PEAK	SEEDS
MEDICAL IMAGING				
VerSe'20	Spine-CT segmentation · SAM ViT-B	+1.21 pp Dice	2.1× fewer	4
MURA	Radiographs, binary · ViT proxy	+2.24 pp *	4.2× fewer	4
NCT-CRC-HE	Histopathology, 9-class · ViT proxy	+0.96 pp †	–	3
NATURAL IMAGES (FROM SCRATCH)				
CIFAR-10	10-class · ViT-Ti/4, native 32×32	+3.12 pp *	2.1× fewer	4 (4/4)
CIFAR-100	100-class · ViT-Ti/4, native 32×32	+3.32 pp *	3.2× fewer	4 (4/4)

* Accuracy gain at AERO's early-stopped peak (standard deployment practice); the compute-to-target figures are independent of where you stop. † NCT-CRC: three-seed matched flip-augmentation result; a second matched recipe (heavy regularization, single configuration) gives +1.03 pp in the same direction. CIFAR wins are unanimous across all four seeds (per-seed spread +2.85 to +3.58 pp on CIFAR-10, +3.07 to +3.81 pp on CIFAR-100).

WHY IT'S A FAIR TEST

- **Both sides tuned, not defaults.** AERO and AdamW were each independently hyperparameter-swept on every regime; AERO is compared against AdamW's own swept best.
- **Only the optimizer changes.** Identical data, model, seed, batch size, budget, schedule, and evaluation in every cell.
- **Multi-seed, mostly unanimous.** Four seeds on most tasks (three on NCT-CRC); on CIFAR, every seed wins.

WHY IT MATTERS

- **The sign and direction are consistent.** The exact number depends on the task, but AERO wins in accuracy and compute in every properly-tuned head-to-head so far — pointing to a general training-loop effect, not a dataset-specific artifact.
- **It doesn't fade on the harder task.** Moving from CIFAR-10 to the harder 100-class CIFAR-100, the margin grows (+3.12 → +3.32 pp) and the compute win widens (2.1× → 3.2×).

SCOPE & HONEST LIMITS

- **Proxy-scale models** — these test breadth across tasks and modalities with off-the-shelf proxies (ViT / ViT-Ti, SAM backbone); model-*scale* is tested separately (a measured 2× token-efficiency result on 1.36B-parameter language pretraining).
- **Supervised training** — fine-tuning and from-scratch classification; RL / post-training remains an expected but unmeasured extension.
- **Public benchmarks** — openly licensed for published benchmarking; not yet specific clinical or production models.
- **Directly measured** — end-to-end on public validation/test splits, mean over independent seeds; not derived or extrapolated.