ECCV 2026 | MALMÖ | SEPT 8–12

Data-Pruning Scores Break Under Distribution Shift

Clean accuracy does not certify a coreset

Vadim Tynchenko · Aleksei Borodulin · Vladislav Kukartsev · Sergei O. Kurashkin

Bauman Moscow State Technical University, Moscow, Russia · kurashkin@bmstu.ru

CDEL: 2nd Workshop on Curated Data for Efficient Learning · ECCV 2026, Malmö

The question

Data pruning scores every training example, keeps the highest-ranked fraction, and trains on that coreset alone. A score is adopted when its subset matches full-data clean accuracy at a budget.

We ask whether clean accuracy certifies the coreset once the test distribution moves.

The benchmark

2
datasets, CIFAR-10 and CIFAR-100

6 + 1
scores against random selection

90–10%
keep ratios, five levels

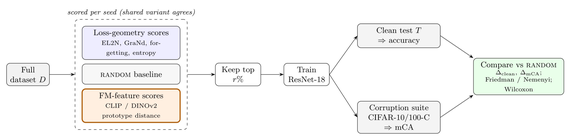
5
seeds per cell

15 × 5
corruptions × severities

360
units, 13.6 GPU-hours

One fixed recipe trains every method: no per-method tuning and no validation-based model selection, so the comparison is an identical budget rather than tuned optima.

Protocol



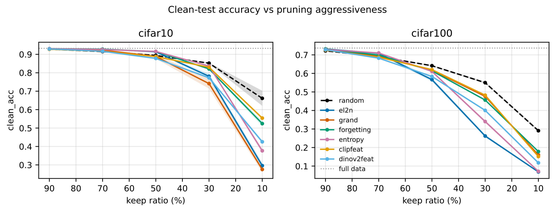
A score orders the full dataset once per seed, the top r% is kept, one ResNet-18 recipe trains on the coreset, and each model is read on the clean test set and on the matched corruption benchmark.

Frozen and gated

Every value comes from one from-scratch run with resume disabled, 360 units behind a review gate, with a per-unit manifest recording the sha256 of the score file each unit consumed.

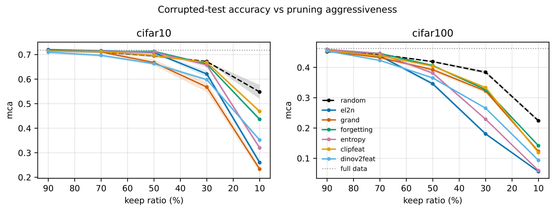
A shared-scoring variant agrees on every sign, rank and significance threshold, so the effect does not depend on one scoring draw. Software versions are pinned per unit and the data checksums are recorded.

Clean accuracy converges



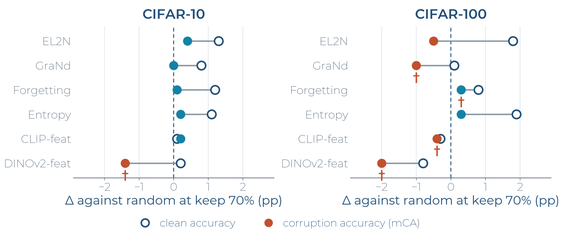
Clean test accuracy against keep ratio. By moderate budgets the curves have merged: most scores reach the bar the standard protocol sets.

Corruption accuracy does not



Mean corruption accuracy for the same models. The curves stay separated where the clean curves merged, several scores sit below random, and below about half the data random is best on both datasets.

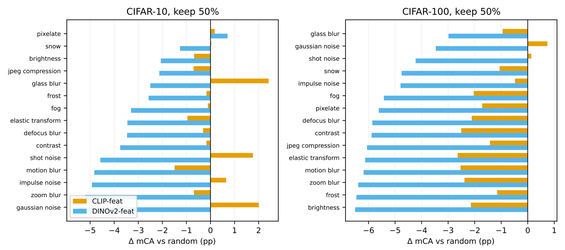
Clean accuracy is not a certificate



Change against random at keep 70%, in percentage points. A dagger marks a Wilcoxon difference from random over the fifteen corruption blocks at $p < 0.05$. DINOv2-feat on CIFAR-10 is +0.2 on clean accuracy, indistinguishable from a random subset, and −1.4 under corruption ($p < 0.001$).

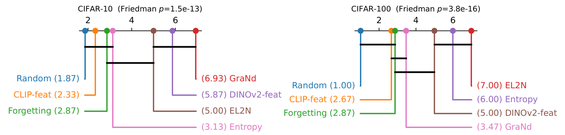
At keep 90% on CIFAR-10 five of six scores beat random on clean accuracy; under corruption only one does. At keep 50% on CIFAR-100 none of the six beats random on either metric.

The two embeddings differ



Change against random per corruption at keep 50%. CLIP-feature pruning is near zero and mixed in sign; DINOv2-feature pruning is negative on almost every corruption.

Ranks under corruption



Nemenyi critical-difference diagram at keep 30%. Random ranks first on CIFAR-100 and second on CIFAR-10, and score identity affects corruption accuracy at every operating point (Friedman $p < 5e{-}4$).

Below half the data, random wins

At keep 30% and 10% the random baseline is best on both datasets, and every score loses at the Wilcoxon floor of p = 6.1e−5 over fifteen corruption blocks. The worst collapse is GraNd on CIFAR-10 at keep 10%: 0.233 mCA against 0.548 for random.

The crossover is dataset-dependent and arrives earlier on the harder dataset, at or above keep 50% on CIFAR-100 and near keep 50% on CIFAR-10.

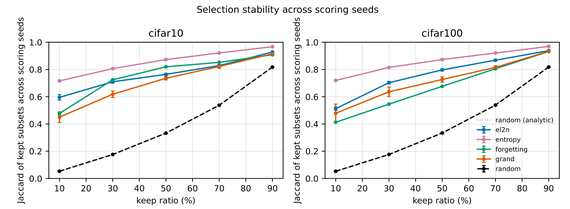
It starves the easy classes

At keep 10% on CIFAR-100 a random coreset keeps about 34 examples of its thinnest class, EL2N keeps 4.8, GraNd 3.0 and DINOv2-feat 2.0.

Keep-hardest scoring removes the easy classes that carry much of the clean accuracy, and worst-class corrupted accuracy collapses with them.

Class bias and corruption fragility appear together on these runs: both are instances of average clean accuracy hiding a cost, and both leave random competitive.

Stable, and stably wrong



Jaccard of the kept subsets across scoring seeds. At keep 50% on CIFAR-10 it is 0.87 for Entropy and 0.76 for EL2N against 0.33 for random: the methods reliably agree on the wrong examples to keep.

Tuning the epoch fails

A self-contained ablation at keep 50%, 110 units, varies one factor at a time. On CIFAR-100 every loss-variant loses to random at $p \le 0.0002$, the best being EL2N at epoch 20 with −4.0 points.

The best epoch inverts across datasets: GraNd at epoch 5 is the best variant on CIFAR-10 and the worst on CIFAR-100, and tuning it would need corrupted validation data a practitioner does not have.

Not a calibration problem

At keep 50% on CIFAR-100 an EL2N coreset scored at epoch 20 is better calibrated than random (clean ECE 0.10 against 0.14) while less accurate and less robust (mCA 0.38 against 0.42). Hard-example selection cures over-confidence without curing fragility.

What we recommend

— Validate a coreset against random selection on a shifted test set, not on the clean one alone.

— Treat CLIP and DINOv2 features as different priors: the embedding governs whether the coreset is robust.

— Release score files and coreset indices so a selection can be re-checked.

Limitations: two datasets at 32 × 32, one architecture and schedule, one shift family of fifteen synthetic corruptions, and a single selection pass. The keep ratio at which random overtakes the scores is a property of this setting.

Code, score files, coreset indices and analysis:
doi.org/10.5281/zenodo.21915235

code and data