Accepted at NeurIPS 2026

Full Fine-Tuning is not the Problem:

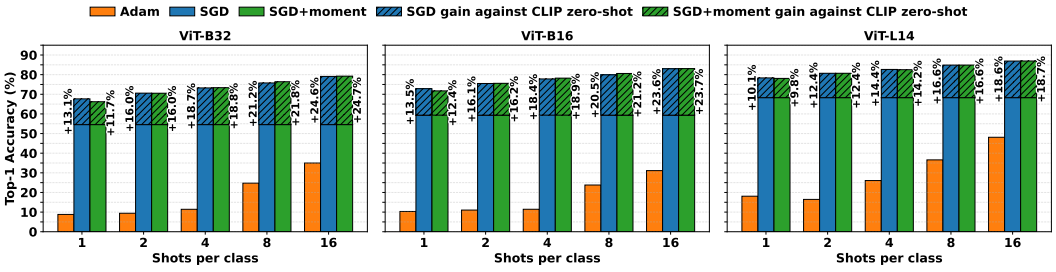
Why Adam Fails and SGD Succeeds in Few-Shot CLIP Adaptation

Nairouz Mrabah Ismail Ben Ayed Eric Granger

LIVIA, ILLS, Department of Systems Engineering
ETS Montreal, Canada

https://github.com/nairouz/few-shot-clip-sgd

Dense adaptation succeeds with plain SGD



The same dense parameterization

SGD substantially exceeds Adam under the shared training recipe.

The central question

Does failure arise from full fine-tuning or from its optimization?

Paper: Fig. 1 and Tables 45–47 (pp. 2, 38–39). Averages over 11 datasets and 3 seeds. Hatched labels are relative gains over zero-shot CLIP.

A controlled benchmark separates these effects

Few-shot image classification
$K \in \{1, 2, 4, 8, 16\}$ labeled examples per class.

11 datasets. Three CLIP backbones: ViT-B/32, ViT-B/16, ViT-L/14.

Three fixed seeds in every setting.

Dense optimizer comparison
30 epochs, batch size 64, learning rate 10^{-4}.

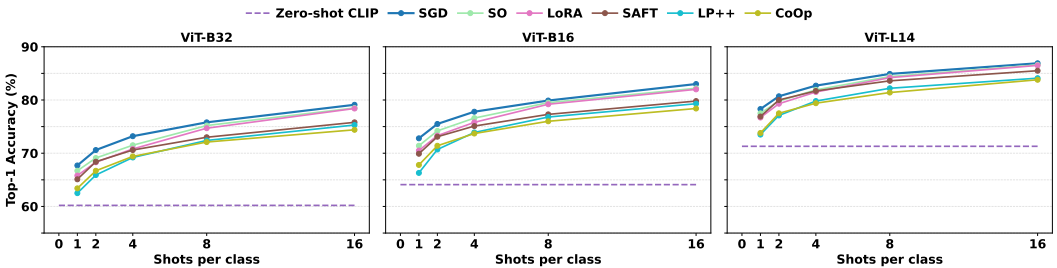
Both encoders are adapted. The common recipe is fixed across datasets, backbones, and shot levels.

PEFT comparisons use fixed method-specific configurations.

The benchmark evaluates specified training recipes, not every possible optimizer tuning.

Paper: Introduction and Appendices B–E, G–H (pp. 2, 30–35, 38–44).

SGD also exceeds the evaluated PEFT averages



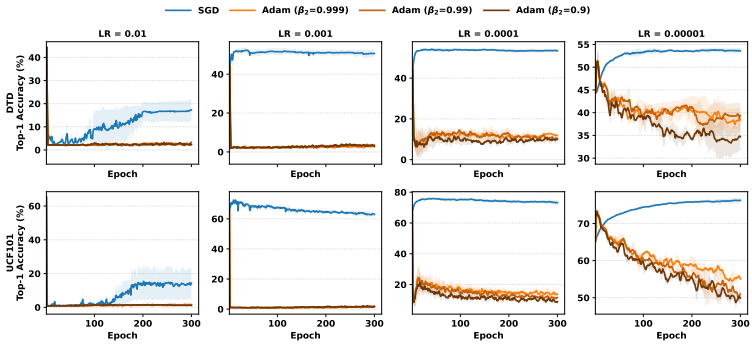
The comparison includes SO, LoRA, SAFT, LP++, and CoOp.

SGD leads the reported dataset-averaged results across backbones and shot levels.

Dense fine-tuning remains a competitive few-shot baseline.

Paper: Fig. 2 and Tables 48–50 (pp. 2, 40–42). This statement concerns the reported averages and protocol, not every dataset or every PEFT recipe.

Changing Adam's second-moment decay is insufficient



SGD exceeds the tested Adam variants across this learning-rate sweep.

One shot. ViT-B/16. DTD and UCF101. Three seeds. Up to 300 epochs.

Paper: Fig. 3 (p. 3). Adam uses $\beta_2 \in \{0.999, 0.99, 0.9\}$. The result is specific to these rates, datasets, and schedules.

Diagnostics compare support and full-data statistics

A smaller mechanistic study

ViT-B/16, one shot, 30 epochs.
EuroSAT, Oxford-Pets, and DTD.
All Transformer blocks in both encoders are trainable.
One support batch per update. No augmentation.

At a fixed checkpoint Θ_t

For encoder e, mean gradients and diagonal second moments are

$$m_{\mathcal{D},t}^{(e)} \in \mathbb{R}^{d_e}, \qquad v_{\mathcal{D},t}^{(e)} \in \mathbb{R}_{\geq 0}^{d_e}.$$

$\mathcal{D}$ is the support or full training split.
d_e is the number of trainable encoder parameters.

These are checkpoint diagnostics. Adam's running moment state is examined separately in Appendix K.4. Full-data statistics are diagnostic references.

Paper: Section 3 and Appendix K (pp. 4–6, 55–85).

The repaired direction retains the support-set numerator

Few-shot direction

$$u_{ff,t}^{(e)} = m_{\mathrm{few},t}^{(e)} \oslash \sqrt{v_{\mathrm{few},t}^{(e)} + \varepsilon}.$$

Both the numerator and denominator come from the support set.

Repaired denominator

$$u_{fr,t}^{(e)} = m_{\mathrm{few},t}^{(e)} \oslash \sqrt{v_{\mathrm{full},t}^{(e)} + \varepsilon}.$$

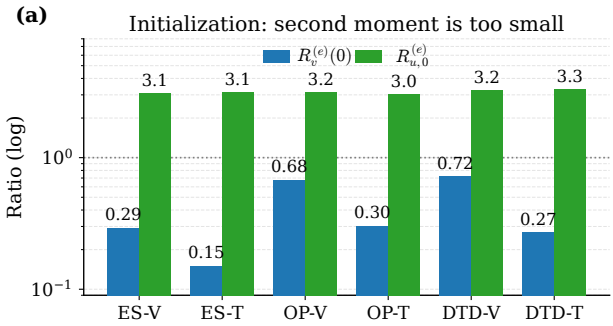
The numerator is unchanged. Only the second moment is replaced.

The full-data reference is $u_{rr,t}^{(e)} = m_{\mathrm{full},t}^{(e)} \oslash \sqrt{v_{\mathrm{full},t}^{(e)} + \varepsilon}$.

All directions lie in $\mathbb{R}^{d_e}$. Division and square root are elementwise. The stabilizer satisfies $\varepsilon > 0$.

Paper: Eq. (3) and Appendix K.5 (pp. 5, 69–71). These counterfactual directions are diagnostic controls.

Miscalibration is present before the first update



$R_v^{(e)}(0)$ compares support and full-data second-moment norms.

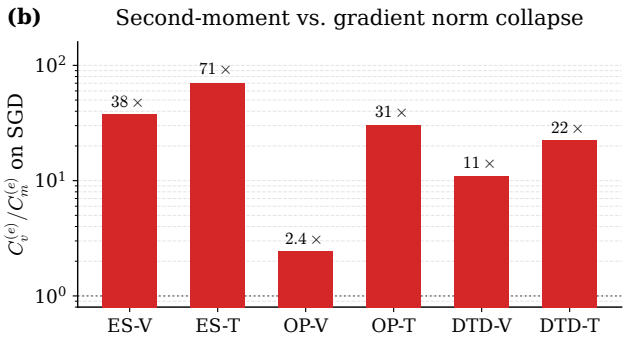
It is below one in all six dataset–encoder pairs.

$R_{u,0}^{(e)}$ compares the norms of $u_{ff,0}^{(e)}$ and $u_{rr,0}^{(e)}$. It is about $3\times$.

ES: EuroSAT. OP: Oxford-Pets. V: vision encoder. T: text encoder.

Paper: Fig. 4(a), Eq. (4), and Figs. 18–20 (pp. 5, 60–61). The adaptive norm ratio changes both numerator and denominator relative to the full-data reference.

The second-moment norm ratio declines more during fitting



On the stable SGD trajectory,

$$\frac{C_v^{(e)}}{C_m^{(e)}} = \frac{R_v^{(e)}(0)/R_v^{(e)}(T)}{R_m^{(e)}(0)/R_m^{(e)}(T)}.$$

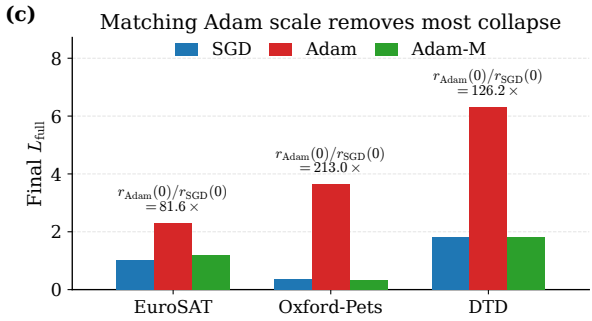
$R_m^{(e)}(t)$ and $R_v^{(e)}(t)$ are support/full norm ratios for gradients and second moments.

The relative decline is larger for the second moment in every pair.

These collapse factors do not by themselves determine the adaptive update norm.

Paper: Fig. 4(b), Eqs. (5)–(6), and Fig. 21 (pp. 5–6, 63). T is the final logged checkpoint. ES/OP and V/T have the meanings on slide M08.

Matching the first update removes most loss collapse



The relative update budget is

$$r_o(t) = \frac{\|\Delta\Theta_t^{(o)}\|_2}{\|\Theta_t^{(o)}\|_2}.$$

Adam-M uses one learning-rate multiplier so that

$$r_{\text{Adam-M}}(0) = r_{\text{SGD}}(0).$$

Adam's adaptive direction and moment recursion are retained.

Early update scale is a direct causal control in this experiment.

Paper: Fig. 4(c), Eq. (7), and Figs. 31–34 (pp. 6, 73–74). $\Delta\Theta_t^{(o)} = \Theta_{t+1}^{(o)} - \Theta_t^{(o)}$. Adam-M is Adam-M0 in Appendix N.

The accuracy recovery extends across all five shot levels

Shots K	SGD	Adam	Adam-M0
1	66.37	6.30	69.28
2	72.28	7.35	72.95
4	75.24	10.64	76.25
8	78.53	18.25	79.18
16	82.23	29.98	82.49
Mean	**74.93**	**14.51**	**76.03**

Accuracy (%). Six CLIP datasets, three seeds, 30 epochs.

The one-time correction recovers the average SGD–Adam gap.

Paper: Table 63 (p. 128). Equal weight per dataset–shot–seed setting. Adam-M0 exceeds SGD by 1.10 points, with paired dataset-cluster 95% interval [0.33, 1.96].

Smaller learning rates and warmup also restore Adam

Optimizer and schedule	Accuracy (%)
SGD, 10^{-4}	74.61
Adam, 10^{-4}	15.04
Adam, 10^{-6}	76.10
Adam, 5% warmup + cosine	76.28
Adam, 20% warmup + cosine	76.26

Six datasets, $K \in \{1, 4, 16\}$, three seeds, 30 epochs.

Warmup uses base rate 10^{-4}. No rate or checkpoint is selected from the test trajectory.

The evidence concerns scale-sensitive failure under a shared recipe.

Paper: selected rows of Table 69 (p. 132). The supplementary slides retain all tested rates and schedules.

Fixed-horizon recovery differs from long-horizon stability

Setting	Epoch 30	Epoch 300	Last	Peak-to-last
Adam, 10^{-6}	79.70	73.29	69.95	10.93
Adam, 10^{-5}	64.71	47.30	46.96	30.21
SGD, 10^{-5}	79.66	80.05	80.05	1.02

One shot. Six datasets. ViT-B/16. Eighteen trajectories per setting.

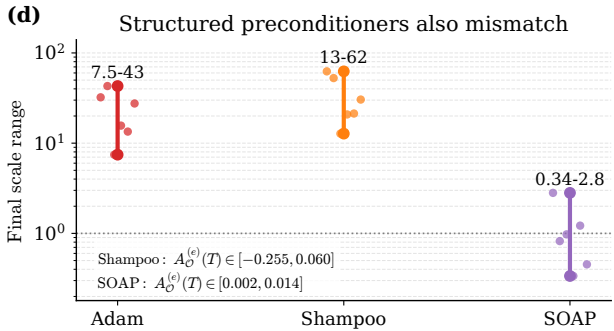
Adam at 10^{-6} has last-observed epochs from 338 to 600.

At 10^{-5}, one Adam run ends at epoch 281.

A lower rate can delay degradation without removing it.

Paper: selected rows of Table 70 and Fig. 70 (pp. 133–134). Accuracies are percentages. Drops are mean within-run differences in percentage points.

Richer preconditioning does not resolve the mismatch



Shampoo

Large scale distortion and poor full-data reference alignment.

SOAP

Smaller scale inflation in several cases, but the repaired direction remains nearly orthogonal to the reference.

$S_{\mathcal{O}}^{(e)}(t) = \|q_{ff,t}^{(\mathcal{O},e)}\|_2 / \|q_{fr,t}^{(\mathcal{O},e)}\|_2$ measures scale. $A_{\mathcal{O}}^{(e)}(t) = \cos(q_{fr,t}^{(\mathcal{O},e)}, q_{rr,t}^{(\mathcal{O},e)})$ compares two directions under the same full-data map.

Paper: Fig. 4(d), Eqs. (8)–(10), and Figs. 43–46 (pp. 6, 82–85). Here $\mathcal{O} \in \{\mathrm{Shampoo}, \mathrm{SOAP}\}$. The cosine measures residual numerator mismatch.

Theory quantifies finite-support estimation error

Proposition 1: a finite-population identity

$$\mathbb{E}\left[\left\|\widehat{v}^{(e)}_{\mathrm{few},t} - v^{(e)}_{\mathrm{full},t}\right\|_2^2\right] = \frac{N-n}{Nn}\sum_{j=1}^{d_e}\sigma^{2,(e)}_{t,j}.$$

Assumptions

The checkpoint is fixed independently of the support draw.

The support contains n uniformly sampled examples without replacement from N, with $1 \le n \le N$, $N > 1$.

Interpretation

$\sigma^{2,(e)}_{t,j}$ is the finite-population variance of squared per-example gradients.

Error depends on gradient variability and support size. Dimension alone does not establish failure.

Paper: Proposition 1, Eqs. (90)–(98) (pp. 85–88). The estimator is unbiased under this idealized model. Experimental support sampling is class-balanced.

Inverse-square-root normalization amplifies uncertainty

Proposition 2: convexity induces upward bias

$$\phi_\varepsilon(x) = (x+\varepsilon)^{-1/2}, \quad x \geq 0,\ \varepsilon > 0,$$
$$\mathbb{E}\Big[\phi_\varepsilon(\widehat{v}^{(e)}_{\mathrm{few},t,j})\Big] \geq \phi_\varepsilon(v^{(e)}_{\mathrm{full},t,j}).$$

The sampling assumptions are those of Proposition 1. The inequality is strict when the estimator has nonzero variance.

Propositions 3–4: alignment alone does not certify a safe step

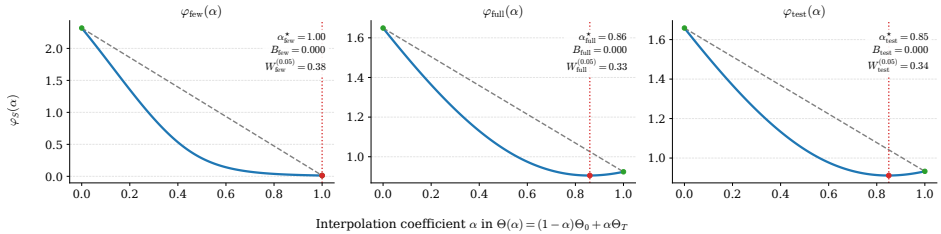
Under twice-continuous differentiability and positive curvature bounds along a fixed direction, a sufficiently long step can increase full-train loss despite positive gradient alignment.

The theory explains a mechanism under stated assumptions. It is not a universal failure guarantee.

Paper: Proposition 2 (pp. 88–90) and Propositions 3–4 (pp. 91–92).

SGD reaches a broad low-loss span along the tested line

$$\Theta(\alpha) = (1-\alpha)\Theta_0 + \alpha\Theta_T, \qquad \alpha \in [0,1].$$



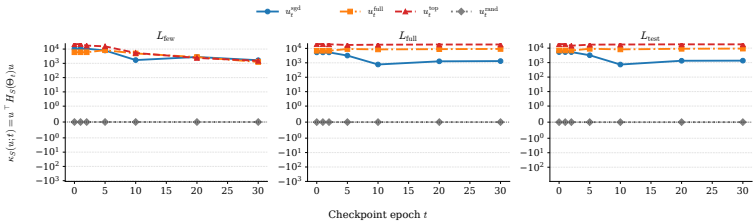
EuroSAT and Oxford-Pets have zero measured full-train and test barriers.

DTD's full-train and test minima occur before the final point.

Paper: Fig. 5(a), Fig. 47, and Appendix M.1 (pp. 7–8, 98–99). Θ_0 is the pretrained checkpoint and Θ_T is the final SGD checkpoint. Broadness is directional.

SGD follows lower positive curvature in the reported probes

$$\kappa_S(u;t) = u^\top H_S(\Theta_t)u, \qquad \|u\|_2 = 1.$$

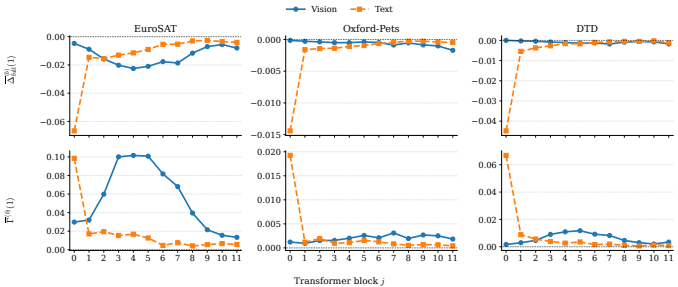


EuroSAT at $t = 30$: full-loss curvature is 1,317 along SGD,

9,158 along the full gradient, and 18,634 along the dominant Hessian direction.

Paper: Fig. 6(a), Fig. 53, Section 4 and Appendix M.4 (pp. 8–9, 106–108). $H_S = \nabla^2 L_S$, $S \in \{\text{few}, \text{full}, \text{test}\}$. The dominant direction is the largest-magnitude full-loss Hessian eigendirection. Comparisons depend on dataset and split.

The useful update budget depends on block and time



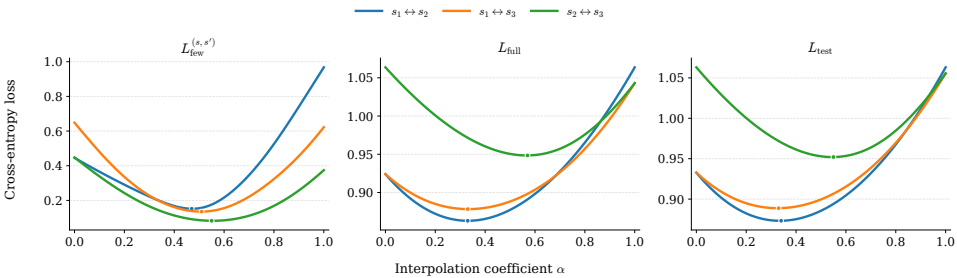
Negative full-loss change indicates beneficial block motion.

Positive mismatch indicates that support loss overstates the benefit.

Many early steps are conservative. Some late directions become support-specific.

Paper: Fig. 6(b) and Figs. 55–57 (pp. 8–9, 109–112). Means over checkpoints $\{0, 1, 2, 5, 10, 20, 30\}$, one-shot ViT-B/16 diagnostics.

SGD solutions are connected across the tested seeds

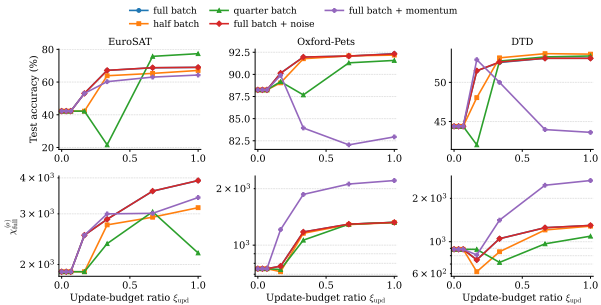


Measured full-train and test barriers are zero for all tested seed pairs on EuroSAT, Oxford-Pets, and DTD.

The endpoints are distinct: relative separation lies between 1.05 and 1.20.

Paper: Fig. 5(b) and Fig. 58 (pp. 8–9, 112–113). The conclusion applies to the sampled straight interpolation segments and grid.

Stochasticity effects depend on the dataset and budget



With matched update counts, quarter-batch SGD improves EuroSAT from 68.98% to 77.36%. Oxford-Pets shows no benefit.

Epoch-, example-, and joint-budget controls are reported in the supplementary slides.

Paper: Fig. 67 and Appendix M.9 (pp. 123–127). Small isotropic noise does not reproduce the gain, but noise magnitude and covariance are not isolated.

A SigLIP replication supports the update-scale explanation

Model	SGD	Adam	Adam-M0
SigLIP	76.28	38.88	76.60

Scope

SigLIP-base-patch16-224.
Four datasets, $K \in \{1, 4, 16\}$.
Three seeds, 30 epochs.
Transformer blocks of both towers are trainable.

Paired comparison

SGD exceeds raw Adam in all 12 dataset–shot settings.
Adam-M0 minus SGD: 0.32 points.
Paired dataset-cluster 95% interval: $[-0.28, 1.29]$.

Paper: Tables 71–72 and Appendix N.7 (pp. 134–135). Values are mean accuracy (%). Embeddings, projection/pooling, post-tower modules, logit scale and bias remain fixed.

The conclusions remain bounded by the evidence

Experimental scope

The main comparison uses fixed recipes and selected PEFT methods.

Adam can recover with scale correction, a smaller learning rate, or warmup.

Fixed-horizon recovery does not imply stability across training time.

Geometric and theoretical scope

Interpolation broadness concerns tested lines. Curvature probes concern selected directions.

Late support specialization remains possible, particularly on DTD.

The finite-population theory assumes an independent checkpoint and uniform support sampling.

Paper: Sections 3–5, Appendices L, M.8–M.9, N.5–N.7 (pp. 6–9, 86–97, 120–135).

Optimization changes the conclusion about full fine-tuning

Dense SGD is effective under the reported few-shot protocol.

Adam's support-set statistics can induce excessive early update scales.
The matched-budget and scheduling controls restore much of the lost accuracy. SGD avoids adaptive preconditioning and exhibits favorable local geometry along the tested directions, with limits from late support specialization.

Optimizer-statistic representativeness is part of the adaptation problem.

`https://github.com/nairouz/few-shot-clip-sgd`

Paper: Abstract, Sections 3–5, and Appendix N.