SPECTRAL GRADIENT MERGING FOR STABLE FEW-SHOT ADAPTATION OF VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot adaptation of vision–language models remains sensitive to the learning rate and training duration, even when SGD achieves strong accuracy. We propose OLM-SGD, an optimizer that combines class-unique minibatching with orthogonal low-rank class-specific gradient merging. Each minibatch contains at most one example per class. For each parameter tensor, OLM-SGD retains dominant per-example singular components and merges them through polar reconstruction. It directly adapts the original encoder blocks without additional trainable modules or optimizer moments. Our analysis establishes retained-gradient alignment and energy control, with conditional stationarity guarantees and finite-shot population bounds that account for support-set reuse. Experiments span 11 datasets, three CLIP backbones, and five shot levels. OLM-SGD improves dataset-average accuracy over the strongest evaluated parameter-efficient or frozen-feature baseline in all 15 backbone–shot settings, with gains up to 1.9 percentage points. One-shot experiments over 300 epochs and four learning rates also show reduced learning-rate sensitivity and improved accuracy retention. At learning rate 10^{-3}, the mean peak-to-final accuracy drop on ViT-B/16 decreases from 6.49 percentage points with SGD to 2.22 with OLM-SGD. Ablations show that polar reconstruction contributes beyond class-unique sampling and low-rank truncation alone. These results support class-specific gradient merging as an effective approach to accurate and stable few-shot adaptation. Our code is included in the supplementary materials and will be released upon publication.

1 INTRODUCTION

Contrastive vision–language pretraining enables CLIP (Radford et al., 2021) to recognize classes from natural-language descriptions without task-specific training. Few-shot adaptation uses a small labeled support set to improve recognition on a target dataset. Existing approaches adapt prompts, add lightweight modules, or fit classifiers over pretrained features. CoOp (Zhou et al., 2022) learns textual contexts, while KgCoOp (Yao et al., 2023) regularizes their embeddings using pretrained knowledge. GDA (Wang et al., 2024) constructs a classifier from frozen support features. These approaches provide different ways to use limited supervision. For adaptation to be reliable, however, strong accuracy should not depend on finding one favorable learning rate and stopping time.

Hyperparameter sensitivity is particularly consequential when labels are scarce. Selecting a configuration on a large target validation set introduces supervision beyond the stated support budget. CLAP (Silva-Rodríguez et al., 2024) examines this evaluation issue and develops a class-adaptive regularized probe. Optimization is also an established concern in prompt learning: ProGrad (Zhu et al., 2023) constrains conflicting prompt updates using pretrained knowledge. More recently, ProLIP (Fahes et al., 2026) combines projector fine-tuning with regularization for reliable adaptation. These results motivate a complementary question: can direct adaptation of the original encoder weights achieve strong performance over a broader range of learning rates and training durations?

The remaining question concerns how the support gradients are combined. SGD averages them before updating the encoder weights. This is the gradient of the minibatch objective, but it no longer exposes the individual class contributions for separate processing. Can controlling the magnitude of this average suffice, or does reliable adaptation also require changing how these contributions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

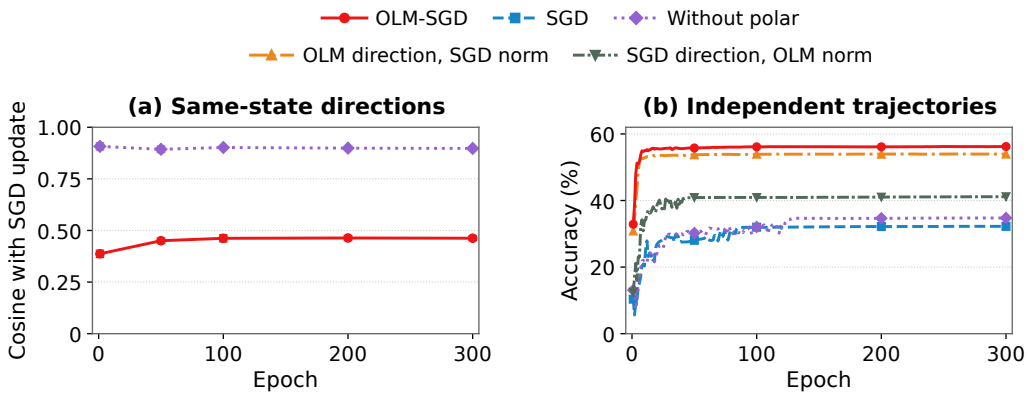

Figure 1: **Gradient geometry and trained interventions.** One-shot ViT-B/16, rank $r = 3$, $\eta = 10^{-3}$, without augmentation. (a) Candidate updates on identical batches along SGD trajectories: DTD, EuroSAT, Aircraft, Cars, and SUN397; three seeds. Points average datasets within each seed, then seeds. Error bars show one sample standard deviation. (b) Independently trained controls on matched completed runs: EuroSAT (three seeds), Aircraft (two), and Cars (one). Curves average seeds within datasets, then weight datasets equally. Norm swaps use the global update norm at each control's own iterate. The two panels use separate experimental campaigns.

determine its direction? Answering this question calls for both diagnostics at a shared parameter state and interventions that change the subsequent training trajectory.

We first compare candidate updates on identical minibatches along SGD trajectories. Each batch contains distinct classes. The candidates are the ordinary gradient mean, the mean of rank-three per-example gradient approximations, and a merge that applies polar reconstruction to their concatenated singular factors. Figure 1(a) reports global cosine similarity to the SGD update on five datasets. Across these probes, truncation retains 83.4% of the individual-gradient energy and yields a cosine of 0.90. Adding polar reconstruction reduces this cosine to 0.45. Thus, polar reconstruction makes a substantial directional change beyond low-rank truncation.

We then train the corresponding interventions independently (Figure 1(b)). Two controls exchange the global norm and direction of SGD and the polar-merged update, recomputed at each control's own parameters. Across the three datasets with matched completed runs at 10^{-3}, the full merge reaches 56.2% mean accuracy at epoch 300. Preserving the merged direction with SGD's norm reaches 53.9%, whereas preserving SGD's direction with the merged update's norm reaches 41.1%. Truncation alone reaches 34.8%. These comparisons support a contribution from the merged direction that scalar norm control does not reproduce. Together, the probes and trained controls motivate processing individual class-specific gradients before merging them, with truncation and polar reconstruction serving distinct roles.

We introduce **OLM-SGD**, an optimizer based on *orthogonal low-rank gradient merging*. Class-unique minibatching selects at most one support example per class in each update. At the same parameter values, OLM-SGD computes the individual-example gradients and decomposes each trainable tensor separately. It retains the leading singular components, concatenates their left and right factors, and replaces these collections with their rectangular polar factors. Reconstruction keeps the retained singular values as coefficients and normalizes by the realized batch size. Each shot therefore contributes through its own gradient rather than a class-wide average. OLM-SGD directly adapts the original residual blocks of both encoders without additional trainable modules or optimizer moments.

The central optimization choice is to preserve class-distinct information until the update is constructed. This differs from restricting adaptation to prompts or modules, and from applying low-rank processing after gradient averaging, as in GaLore (Zhao et al., 2024). The singular-factor transformation builds on TSV-M (Gargiulo et al., 2025), which merges independently trained task vectors. Here, class-unique sampling supplies instantaneous gradients at one shared iterate. Each merge becomes an optimization step and changes the gradients used by the next step. Our analysis estab-

lishes alignment with the retained gradient mean and an update-norm bound determined by retained energy. It further gives conditional stationarity guarantees and finite-shot population bounds that account for reuse of the fixed support set.

The broader evaluation tests whether these properties translate into accurate and stable adaptation. Across 11 datasets, three CLIP backbones, and five shot levels, OLM-SGD achieves the highest dataset-average accuracy among the evaluated parameter-efficient and frozen-feature baselines in all 15 backbone–shot settings. Gains reach 1.9 percentage points. One-shot trajectories over 300 epochs and four learning rates also show improved accuracy retention. On one-shot ViT-B/16 at 10^{-3}, the 11-dataset mean per-run peak-to-final drop decreases from 6.49 points with SGD to 2.22 with OLM-SGD. These results connect the gradient-merging construction to both few-shot accuracy and reduced dependence on the stopping time.

Contributions. **(1)** Same-state probes and matched training interventions distinguish the effect of polar reconstruction from truncation and scalar norm control. **(2)** OLM-SGD combines class-unique sampling with tensor-wise merging of individual gradients to adapt the original encoder blocks. **(3)** Retained-gradient alignment and energy bounds characterize the update, with conditional stationarity and finite-shot population analyses. **(4)** Evaluation across 15 backbone–shot settings establishes consistent dataset-average gains over the evaluated parameter-efficient and frozen-feature baselines, complemented by long-horizon and component studies.

2 Related Work

Few-Shot Adaptation of Vision–Language Models. Few-shot CLIP (Radford et al., 2021) adaptation often restricts the trainable components. CoOp (Zhou et al., 2022) learns textual contexts. KgCoOp (Yao et al., 2023) regularizes their embeddings toward handcrafted prompts, while ProGrad (Zhu et al., 2023) constrains prompt gradients using pretrained predictions. CLIP Adapter (Gao et al., 2024) learns residual feature transformations, TIP-Adapter-F (Zhang et al., 2022) tunes support-cache keys, and ProLIP (Fahes et al., 2026) regularizes the visual projection. These methods freeze the transformer blocks of the vision and text encoders. Unlike these methods, OLM-SGD updates both encoders' transformer blocks, allowing supervision to reshape the feature extractor.

LoRA (Hu et al., 2022) learns factorized weight corrections. PiSSA (Meng et al., 2024) trains principal weight components with a frozen residual. DoRA (Liu et al., 2024) combines low-rank directional adaptation with learned magnitudes. SAFT (Nguyen et al., 2025) selects a fixed sparse parameter support, whereas SO (Mrabah et al., 2025) changes its sparse support using adaptive moments. These methods constrain the parameterization or accessible coordinates. OLM-SGD instead controls adaptation through its structured gradient-merging update. More precisely, the proposed optimizer can be used as a full fine-tuning technique in few-shot adaptation. It constructs class-specific individual tensor gradients before merging and directly updates all transformer blocks and their layers. Its retained gradient rank does not impose a fixed-rank cumulative weight change.

GDA (Wang et al., 2024) and ProKeR (Bendou et al., 2025) construct discriminant and proximal-kernel classifiers over frozen features. Their supervision changes the decision rule within an existing representation. Unlike these methods, OLM-SGD is an optimization technique. It also adapts the representation without adding trainable modules. Yet, it controls this broader adaptation through class-distinct gradient geometry.

Model Merging. Task Arithmetic (Ilharco et al., 2023) represents downstream adaptations as parameter differences from a shared pretrained model and combines these task vectors through arithmetic operations. TSV-M (Gargiulo et al., 2025) further targets interference among task vectors by retaining dominant singular components and orthogonalizing their concatenated singular subspaces before reconstruction. OLM-SGD adopts this spectral principle in a fundamentally different setting. TSV-M performs *offline* merging of models obtained from separately trained tasks. For TSV-M, each dataset represents a task. In contrast, OLM-SGD applies low-rank orthogonal merging *during optimization* to gradients computed at the current iterate. It aims to reduce interference between classes. Class-unique minibatching further ensures that the merged gradients correspond to distinct classes, thereby directing the spectral transformation toward cross-class interactions. The resulting merge is the optimization update itself. It changes the model parameters and therefore changes the gradients observed at subsequent iterations. Thus, OLM-SGD converts a static model-merging con-

struction into an online, state-dependent optimization rule. Our analysis further characterizes this rule through the retained gradient energy and its alignment with the population gradient.

Structured Optimization. SGD averages minibatch gradients. Adam (Kingma & Ba, 2015) rescales them using historical moments. GaLore (Zhao et al., 2024) projects aggregated gradients into reduced coordinates. Aligned-MTL (Senushkin et al., 2023) conditions a matrix of vectorized task gradients, while Muon transforms matrix momentum through an approximate polar operation (Jordan, 2024). OLM-SGD decomposes individual tensor gradients before merging their factors, without persistent moments. It therefore uses information no longer separately accessible from the minibatch mean. Two-sided factor transformations modify direction and scale while retaining individual singular values as reconstruction coefficients.

3 OLM-SGD: Orthogonal Low-Rank Gradient Merging

OLM-SGD is a full fine-tuning optimization technique built on two complementary components. *Class-unique minibatching* ensures that each minibatch contains at most one support example per class, so the per-example gradients merged in an update originate from distinct classes. *Orthogonal low-rank merging* then retains their dominant tensor-wise components and orthogonalizes the resulting subspaces before reconstructing the update.

3.1 Class-Unique Sampling

Let $\mathcal{D}_{\text{sup}} = \{(x_i, y_i)\}_{i=1}^{N}$ be a few-shot support set with K examples from each of C classes, so that $N = CK$. Let $\Theta \in \mathbb{R}^d$ denote the vector of trainable parameters, initialized from the pretrained CLIP parameters Θ_0, where d is the number of trainable scalar parameters. All remaining model parameters are frozen. We minimize the empirical support objective

$$F(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i(\Theta), \tag{1}$$

where $\ell_i(\Theta)$ denotes the CLIP classification loss for the support example (x_i, y_i). Appendix A describes the CLIP few-shot adaptation setting and provides the explicit form of this loss. Class-unique sampling changes only how support examples are grouped into minibatches. Each loss ℓ_i remains a C-class classification loss and is computed using the logits of all C classes.

We construct minibatches with distinct class labels. Let $1 \le B \le C$ be the maximum minibatch size, and let $\mathcal{I}_c$ contain the K support indices of class c. At the beginning of each epoch, we independently shuffle every $\mathcal{I}_c$. For each round $k \in \{1, \dots, K\}$, we sample a fresh uniform permutation π_k of the C classes and partition it into groups of at most B classes. For each group, we form a minibatch by selecting the kth shuffled example from every class in that group. The final group is retained if it contains fewer than B classes.

Hence, every minibatch $\mathcal{B}_t$ has the size $B_t \le B$ and contains at most one example from each class. Each support example is used exactly once per epoch, resulting in $K\lceil C/B \rceil$ parameter updates. For $K > 1$, different examples from the same class appear in different minibatches. Algorithm 1 provides the complete Class-Unique minibatch sampling technique.

Let the support indices in $\mathcal{B}_t$ be $i_{t,1}, \dots, i_{t,B_t}$. Before updating the parameters, we evaluate all per-example gradients at the same iterate Θ_t:

$$\mathcal{L}_t(\Theta_t) = \frac{1}{B_t} \sum_{b=1}^{B_t} \ell_{i_{t,b}}(\Theta_t), \qquad g_t^{(b)} = \nabla_\Theta \ell_{i_{t,b}}(\Theta_t) \in \mathbb{R}^d. \tag{2}$$

Ordinary SGD applies the minibatch-mean gradient

$$g_t^{\text{SGD}} = \frac{1}{B_t} \sum_{b=1}^{B_t} g_t^{(b)}. \tag{3}$$

OLM-SGD instead retains the B_t per-example gradients as separate inputs to the merge operator. Because the minibatch is class-unique, these gradients correspond to pairwise distinct class labels.

3.2 ORTHOGONAL LOW-RANK GRADIENT MERGING

Let $\mathcal{W}$ denote the ordered collection of trainable parameter tensors whose entries form Θ. For each $W \in \mathcal{W}$, let $g_t^{(b),W}$ denote the part of $g_t^{(b)}$ associated with W. The map mat_W reshapes this tensor into a matrix without changing its entries. The first tensor axis defines the rows, while all remaining axes are flattened into columns. For a rank budget $r \geq 1$, OLM-SGD defines

$$G_t^{(b),W} = \mathrm{mat}_W(g_t^{(b),W}) \in \mathbb{R}^{m_W \times n_W}, \qquad r_W = \min\{r, m_W, n_W\}. \tag{4}$$

Here, m_W and n_W are the dimensions of $G_t^{(b),W}$, and r_W is the effective rank used for tensor W.

For clarity, we now consider one tensor W. OLM-SGD applies the singular value decomposition (SVD) to every matrix $G_t^{(b)}$ and retains its leading r_W singular components:

$$\Pi_{r_W}(G_t^{(b)}) = U_t^{(b)} \Sigma_t^{(b)} V_t^{(b)\top}. \tag{5}$$

Here, $\Pi_{r_W}(G_t^{(b)})$ is the rank-r_W truncated SVD of $G_t^{(b)}$, $U_t^{(b)} \in \mathbb{R}^{m_W \times r_W}$ is the matrix of left singular vectors and $V_t^{(b)} \in \mathbb{R}^{n_W \times r_W}$ is the matrix of right singular vectors. Both have orthonormal columns, while $\Sigma_t^{(b)} \in \mathbb{R}^{r_W \times r_W}$ is a diagonal matrix. Its entries satisfy $\sigma_{t,1}^{(b)} \geq \cdots \geq \sigma_{t,r_W}^{(b)} \geq 0$.

Then, we concatenate the retained left and right singular factors across the B_t examples

$$U_t = [U_t^{(1)} | \cdots | U_t^{(B_t)}] \in \mathbb{R}^{m_W \times R_t}, \quad V_t = [V_t^{(1)} | \cdots | V_t^{(B_t)}] \in \mathbb{R}^{n_W \times R_t}, \tag{6}$$

$$s_t = \frac{1}{B_t}(\sigma_{t,1}^{(1)}, \ldots, \sigma_{t,r_W}^{(1)}, \ldots, \sigma_{t,1}^{(B_t)}, \ldots, \sigma_{t,r_W}^{(B_t)})^\top \in \mathbb{R}_{\geq 0}^{R_t}. \tag{7}$$

Here, $R_t = r_W B_t$. The vector s_t stacks the retained singular values in the same order as the corresponding columns of U_t and V_t. The factor $1/B_t$ ensures that these factors reconstruct the mean of the truncated per-example gradients:

$$\widetilde{G}_t = U_t \,\mathrm{diag}(s_t) V_t^\top = \frac{1}{B_t} \sum_{b=1}^{B_t} \Pi_{r_W}(G_t^{(b)}) \in \mathbb{R}^{m_W \times n_W}. \tag{8}$$

Thus, $\widetilde{G}_t$ is exactly the arithmetic mean of the rank-r_W gradient approximations.

Although the columns of each $U_t^{(b)}$ and $V_t^{(b)}$ are orthonormal, singular directions from different classes can overlap after concatenation. Consequently, $U_t^\top U_t$ and $V_t^\top V_t$ are not generally identity matrices. Following the singular-factor merging principle of TSV-M (Gargiulo et al., 2025), OLM-SGD reduces this cross-class overlap by replacing each concatenated factor with its polar factor.

For a matrix $A \in \mathbb{R}^{m_A \times n_A}$, let $A = P_A D_A Q_A^\top$ be its reduced SVD, where $q_A = \min(m_A, n_A)$, $P_A \in \mathbb{R}^{m_A \times q_A}$, $D_A \in \mathbb{R}^{q_A \times q_A}$, and $Q_A \in \mathbb{R}^{n_A \times q_A}$. We define

$$\mathrm{polar}(A) = P_A Q_A^\top. \tag{9}$$

For $m_A \geq n_A$, $\mathrm{polar}(A)$ has orthonormal columns. For $m_A < n_A$, it has orthonormal rows. In either case, it is a closest matrix to A in Frobenius norm under the corresponding orthogonality constraint (Higham, 1986). Thus, the transformation modifies the concatenated singular directions as little as possible while enforcing the orthogonality permitted by their dimensions. For rank-deficient A, the closest factor need not be unique, and the reduced SVD provides a valid completion.

The final matrix update for tensor W is

$$\widehat{G}_t = \mathrm{polar}(U_t)\,\mathrm{diag}(s_t)\,\mathrm{polar}(V_t)^\top \in \mathbb{R}^{m_W \times n_W}. \tag{10}$$

OLM-SGD replaces the concatenated left and right singular factors by their polar factors while keeping the scaled singular values s_t as reconstruction coefficients. This operation acts on both sides of the gradient matrix and can change both the direction and magnitude of the retained mean.

Finally, OLM-SGD restores each $\widehat{G}_t^W$ to the shape of its corresponding parameter tensor and concatenates all entries according to the fixed order of $\mathcal{W}$. This gives the merged update $\widehat{g}_t \in \mathbb{R}^d$. The parameter update is

$$\Theta_{t+1} = \Theta_t - \eta_t \widehat{g}_t, \qquad \eta_t > 0, \tag{11}$$

where η_t is the learning rate at iteration t. Algorithm 2 provides the complete procedure. Section 4 analyzes the retained energy of this update and its conditions for optimization progress.

4 Optimization and Few-Shot Population Analysis

We analyze the empirical support objective F defined in equation 1 under the OLM-SGD update in equation 11. We use the notation introduced in Section 3 throughout. Appendix E provides the proofs, conditional update-wise rates, and finite-shot population bounds.

Geometry of the merge. Consider the tensor-wise quantities defined in Section 3.2. When needed, we add the superscript W to identify the trainable tensor to which a quantity belongs. Thus, U_t^W, V_t^W, s_t^W, $\widetilde{G}_t^W$, and $\widehat{G}_t^W$ denote the corresponding quantities for $W \in \mathcal{W}$.

Vectorizing each retained mean $\widetilde{G}_t^W$ and concatenating the entries in the same fixed parameter order used to construct $\widehat{g}_t$ gives $\widetilde{g}_t \in \mathbb{R}^d$. We define the retained-energy budget

$$\mathcal{Q}_t = \sum_{W \in \mathcal{W}} \|s_t^W\|_2^2. \tag{12}$$

We use $\|\cdot\|_2$ for the Euclidean norm of a vector and the spectral norm of a matrix. We use $\|\cdot\|_F$ for the Frobenius norm and $\langle\cdot,\cdot\rangle$ for the Euclidean inner product.

Lemma 4.1 (Retained alignment and energy)**.** *Assume exact singular value and polar decompositions. The following bounds hold for wide or rank-deficient factors:*

$$\langle \widetilde{g}_t, \widehat{g}_t \rangle \geq \sum_{W \in \mathcal{W}} \frac{\|\widetilde{G}_t^W\|_F^2}{\|U_t^W\|_2 \|V_t^W\|_2} \geq B_t^{-1} \|\widetilde{g}_t\|_2^2, \tag{13}$$

$$\|\widehat{g}_t\|_2^2 \leq \mathcal{Q}_t \leq B_t^{-2} \sum_{b=1}^{B_t} \|g_t^{(b)}\|_2^2. \tag{14}$$

The equality $\|\widehat{g}_t\|_2^2 = \mathcal{Q}_t$ holds when both polar factors have orthonormal columns for every tensor.

Thus, polar reconstruction has positive inner product with $\widetilde{g}_t$ whenever $\widetilde{g}_t \neq 0$. A factor is wide when it has more columns than rows. It is rank-deficient when its rank is smaller than the minimum of its two dimensions. Appendix E.3 gives an exact expression for the energy gap $\mathcal{Q}_t - \|\widehat{g}_t\|_2^2$ for these cases. Alignment with the empirical gradient $\nabla F(\Theta_t) \in \mathbb{R}^d$ also depends on truncation and sampling. Appendix E.4 gives sufficient conditions for this alignment.

Convergence with class-unique reshuffling. Let $E \geq 1$ be the number of epochs and $e \in \{1, \ldots, E\}$ the epoch index. Each epoch has $J = K\lceil C/B \rceil$ updates and starts at iteration $t_e = (e-1)J$. Define the epoch-start empirical gradient and the mean merged update as

$$f_e = \nabla F(\Theta_{t_e}) \in \mathbb{R}^d, \qquad h_e = \frac{1}{J} \sum_{t=t_e}^{t_e+J-1} \widehat{g}_t \in \mathbb{R}^d. \tag{15}$$

To isolate the effect of unequal batch sizes, all realized batch gradients should be evaluated at the fixed epoch-start parameters:

$$\omega_e = \frac{1}{J} \sum_{t=t_e}^{t_e+J-1} \frac{1}{B_t} \sum_{b=1}^{B_t} \nabla_\Theta \ell_{i_{t,b}}(\Theta_{t_e}) - f_e \in \mathbb{R}^d. \tag{16}$$

The index $i_{t,b}$ identifies example b of minibatch t, and $\ell_{i_{t,b}} : \mathbb{R}^d \to \mathbb{R}$ is its loss. Thus, ω_e measures the discrepancy between averaging batch means and averaging individual-example gradients at the same parameters.

Let $\mathcal{H}_{t_e}$ denote the training history up to the start of epoch e. It contains the fixed support set, the current iterate Θ_{t_e}, and all training randomness revealed before the new epoch shuffle. Define the nonnegative scalars

$$D_e = \|h_e - f_e - \omega_e\|_2, \qquad \nu_e^2 = \mathbb{E}\left[\|\omega_e\|_2^2 \mid \mathcal{H}_{t_e}\right]. \tag{17}$$

Here, D_e measures the remaining epoch error from merging and within-epoch parameter drift. The quantity ν_e^2 is the conditional mean squared batch-weighting error. Fresh independent shuffles give $\mathbb{E}[\omega_e \mid \mathcal{H}_{t_e}] = 0$. Equal batch sizes give $\omega_e = 0$ for every shuffle. Appendix E.11 bounds D_e using merge distortion and energy-controlled parameter drift. It also gives the exact expression for ν_e^2.

Theorem 4.2 (Quantified epoch stationarity). *Fix the support and use fresh independent epoch shuffles. Let $F_{\inf} \in \mathbb{R}$ be a finite lower bound of F. Assume that ∇F is L_F-Lipschitz on each segment joining an epoch start to its endpoint. Here, $L_F > 0$ is a scalar. Gradient differences have norm at most L_F times the corresponding parameter distance. Use a deterministic scalar learning rate $\eta_e > 0$ throughout epoch e. Define the scalar epoch step $\gamma_e = J\eta_e$ and assume $0 < \gamma_e \le L_F^{-1}$. Let $\Lambda_E^{\mathrm{epoch}} = \sum_{e=1}^{E} \gamma_e > 0$ be the total epoch step. Let $\Delta_F = F(\Theta_0) - F_{\inf} \ge 0$ be the initial objective gap. Both quantities are scalars. If all expectations below are finite, then*

$$\frac{\sum_{e=1}^{E} \gamma_e \mathbb{E}[\|f_e\|_2^2]}{\Lambda_E^{\mathrm{epoch}}} \le \frac{2\Delta_F + \sum_{e=1}^{E} \gamma_e (1 + L_F \gamma_e) \mathbb{E}[D_e^2]}{\Lambda_E^{\mathrm{epoch}}} + \frac{2L_F \sum_{e=1}^{E} \gamma_e^2 \mathbb{E}[\nu_e^2]}{\Lambda_E^{\mathrm{epoch}}}. \tag{18}$$

The left-hand side is a weighted average of squared empirical gradient norms at epoch starts. Equivalently, it is the expected squared gradient norm at an epoch start selected with probability $\gamma_e / \Lambda_E^{\mathrm{epoch}}$. A persistent reconstruction error can leave a nonzero residual in the bound. Corollary E.16 gives a rate under its relative-error condition. This condition allows part of the mean squared reconstruction error to scale with $\|f_e\|_2^2$. Suppose its additive residual is zero and $\mathbb{E}[\nu_e^2]$ is uniformly bounded. Then, admissible epoch steps of order $E^{-1/2}$ yield a stationarity bound of order $E^{-1/2}$. The analysis retains within-epoch dependence and incomplete-batch weighting.

5 Results and Discussion

Protocol. We evaluate few-shot CLIP adaptation on 11 datasets: ImageNet (Deng et al., 2009), SUN397 (Xiao et al., 2010), FGVC-Aircraft (Maji et al., 2013), EuroSAT (Helber et al., 2019), Stanford Cars (Krause et al., 2013), Food101 (Bossard et al., 2014), Oxford Pets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Caltech101 (Fei-Fei et al., 2007), DTD (Cimpoi et al., 2014), and UCF101 (Soomro et al., 2012). We use 3 vision backbones ViT-B/32, ViT-B/16, and ViT-L/14 with $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class and seeds 1, 2, and 3. We report top-1 test accuracy. Dataset averages weight all datasets equally. Aggregate uncertainty is the population standard deviation across seeds after averaging datasets within each seed.

The comparisons cover prompt learning (CoOp (Zhou et al., 2022), KgCoOp (Yao et al., 2023), ProGrad (Zhu et al., 2023)), adapter, cache, and projection tuning (Adapter (Gao et al., 2024), TIP-Adapter-F (Zhang et al., 2022), ProLIP (Fahes et al., 2026)), low-rank adaptation (LoRA (Hu et al., 2022), DoRA (Liu et al., 2024), PiSSA (Meng et al., 2024)), and sparse updates (SAFT (Nguyen et al., 2025), SO (Mrabah et al., 2025)). GDA (Wang et al., 2024) and ProKeR (Bendou et al., 2025) provide frozen-feature alternatives. We also evaluate robust fine-tuning with VRF (Zhu et al., 2024), WiSE (Wortsman et al., 2022b), and CLIPood (Shu et al., 2023) using Adam and SGD. The ablations examine the standard optimizers Adam and SGD, class-unique sampling, scale control, low-rank processing, class-mean gradients, and polar reconstruction. Appendix C provides the ablation variant algorithms, and Appendix D specifies the hyperparameters of all the compared methods.

OLM-SGD updates all the layers of the Transformer residual blocks of both encoders with class-unique batch capacity $B = \min\{64, C\}$ and rank $r = 3$. Its fixed schedule uses 70 epochs at 10^{-4}. Baselines retain their method-specific schedules. Standard comparisons use deterministic resizing and center cropping without data augmentation. Separately, the trajectory diagnostic uses 1-shot ViT-B/16 adaptation over 300 epochs at $\eta \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. Evaluation is deterministic. Appendices F–I cover the comprehensive results: few-shot adaptation baselines, robustness, and epoch transfer in Appendix F; robust fine-tuning in Appendix G; optimizer and component ablations in Appendix H; and rank sensitivity in Appendix I.

Main comparison. *Few-Shot Accuracy across backbones and shots.* Table 1 shows that OLM-SGD attains the highest dataset-average accuracy in all 15 backbone–shot settings. The strongest competitor changes across settings, from SO to PiSSA or GDA. OLM-SGD exceeds these setting-specific leaders by 0.9–1.8 percentage points on ViT-B/32, 1.2–1.9 on ViT-B/16, and 0.1–0.8 on ViT-L/14. Its 16-shot averages reach 78.5%, 82.6%, and 86.8%, respectively. The gains persist across model scales and are larger on the ViT-B backbones. On Aircraft, 16-shot ViT-B/16 accuracy reaches 52.9%, compared with 48.3% for GDA. These results indicate that effective few-shot adaptation does not require restricting the trainable parameter set. OLM-SGD instead controls adaptation

Table 1: Few-shot top-1 accuracy (%), averaged over 11 datasets and three seeds under the method-specific schedules. Columns give shots per class. The selected baselines include every best baseline across the 15 settings. Bold denotes the highest mean and underlining the strongest baseline. Appendix F.1 reports all 13 baselines and per-dataset means and standard deviations.

	ViT-B/32					ViT-B/16					ViT-L/14				
Method	1	2	4	8	16	1	2	4	8	16	1	2	4	8	16
LoRA	65.5	67.6	69.9	73.5	76.6	69.8	72.7	74.7	77.7	80.6	77.0	79.0	81.2	83.6	85.8
SAFT	65.1	68.4	70.6	73.0	75.8	69.9	73.1	75.1	77.3	79.8	76.9	80.0	81.7	83.6	85.5
SO	66.3	68.7	70.7	73.7	75.8	70.8	73.5	75.9	78.4	80.4	77.3	79.4	80.8	82.6	84.0
PiSSA	64.8	68.3	71.3	74.1	76.8	69.7	72.9	76.0	78.4	80.9	77.0	80.0	82.1	84.1	86.0
GDA	42.8	66.9	70.5	73.9	77.0	47.0	71.4	75.1	78.0	81.1	54.4	78.0	81.2	83.4	85.2
OLM-SGD	**67.2**	**70.4**	**73.1**	**75.8**	**78.5**	**72.0**	**75.1**	**77.8**	**80.3**	**82.6**	**77.4**	**80.1**	**82.5**	**84.7**	**86.8**

through its structured gradient-merging update (Appendix F.1).Against robust fine-tuning, OLM-SGD exceeds VRF-SGD, the strongest evaluated variant, in 14 of 15 settings under the reported schedules (Appendix G).

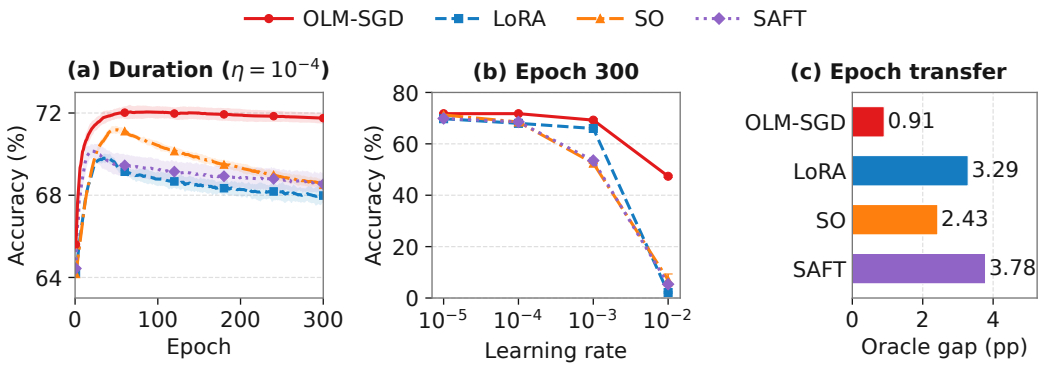

Figure 2: Accuracy retention and duration transfer for 1-shot ViT-B/16 on all 11 datasets, seeds 1–3. (a) Dataset-average accuracy at $\eta = 10^{-4}$. (b) Accuracy at epoch 300 across four learning rates. Bands in (a) and error bars in (b) show one population standard deviation after averaging datasets within each seed. (c) Mean target-oracle gaps over the 110 off-diagonal epoch transfers at 10^{-4}. The source selects the earliest peak of its seed-mean test curve over epochs 1–300. Only the selected epoch is transferred.

Learning rate and prolonged training. Figure 2(a,b) illustrates the evolution of the test accuracy beyond the fixed schedules. At 10^{-4}, OLM-SGD retains 71.75% at epoch 300, versus 67.98–68.54% for LoRA, SO, and SAFT. Its mean per-run peak-to-final drop is 0.95 points, compared with 2.92–3.83 for the best baselines (i.e., LoRA, SO, and SAFT). Across 10^{-3}–10^{-5}, the final-accuracy range for OLM-SGD is 2.52 points, versus 3.79–18.76 for the same baselines. The full grid includes the stress-test rate 10^{-2}, where all methods deteriorate. Strong final accuracy together with smaller degradation supports a wider useful optimization range (Appendix F.3).

Transfer of the selected duration. Figure 2(c) asks how much accuracy is retained when one dataset selects the stopping epoch for an independently trained target. Selection uses source *test* accuracy, so this is a retrospective diagnostic. Across distinct source–target pairs, OLM-SGD obtains 71.73% transfer accuracy and a 0.91-point gap to the target's own best seed-mean checkpoint. SO reaches 69.78% with a 2.43-point gap. OLM-SGD stays within one point of the target oracle on $85/110$ transfers, versus $55/110$ for SO. Its higher transfer accuracy therefore combines a stronger attainable peak with a smaller stopping-time penalty (Appendix F.2).

Ablations. *Optimizer and sampling.* Standard SGD is substantially stronger than Adam under the fixed schedules. OLM-SGD further improves on SGD by 0.3–1.2 points across the ten ViT-B

settings (Appendix H.1). These tables use 30 epochs for SGD and 70 for OLM-SGD. The 30-epoch SGD schedule is consistent with its long-horizon behavior at the same learning rate. At $\eta = 10^{-4}$, the dataset-average accuracy for SGD peaks at epoch 21 with 71.61% and remains at 71.49% at epoch 30, only 0.12 points below the peak. The common 300-epoch diagnostic gives a complementary comparison: at 10^{-3}, OLM-SGD reaches 69.23%, compared with 58.67% for randomized SGD and 59.68% for our variant class-unique SGD. At one shot, both SGD samplers already have the same batching distribution. Therefore, the advantage supports the merging rule beyond class-uniqueness alone (Appendix H.2).

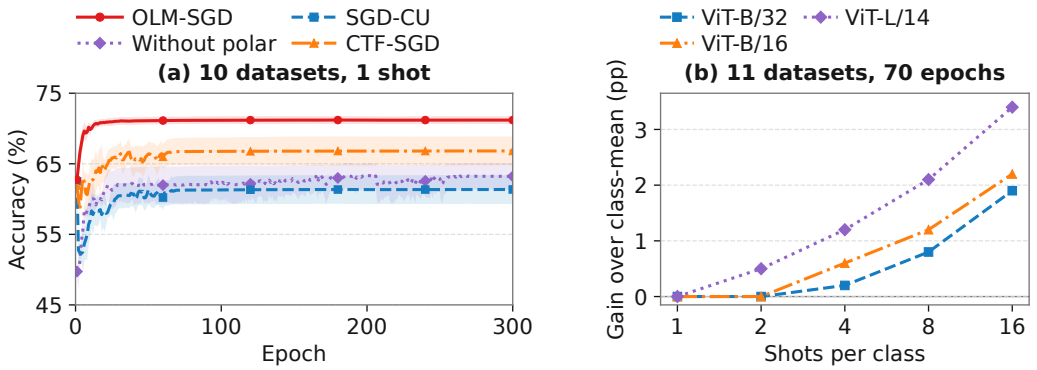

Figure 3: Component comparisons over seeds 1–3. (a) One-shot ViT-B/16 at $\eta = 10^{-3}$, averaged over the ten non-ImageNet datasets with paired polar-ablation histories. Bands show one population standard deviation of the dataset averages. (b) OLM-SGD minus its class-mean variant on all 11 datasets, using the reported fixed-schedule means. Both use 70 epochs, $\eta = 10^{-4}$, and $r = 3$. Equal epochs match example exposure. OLM-SGD takes K times as many updates its class-mean variant.

Scale control and spectral construction. Our ablation variant Cross-Term-Free SGD (CTF-SGD) suppresses excess update magnitude while preserving the SGD direction (Appendix C.2). It reaches 64.67% at 10^{-3}, recovering part of the gap but remaining 4.56 points below OLM-SGD (Appendix H.3). It is a scale-control baseline, not an exact OLM-SGD norm match. GaLoRE instead applies low-rank projection and Adam-style moments after gradient averaging. At 10^{-4}, its final accuracy is 67.78%, versus 71.75% for OLM-SGD (Appendix H.4). These controls do not reproduce the benefit of transforming individual class-specific gradients before merging.

Polar reconstruction and individual shots. Omitting polar reconstruction preserves class-unique sampling and gradient truncation. On the ten paired datasets, restoring it improves every fixed-schedule ViT-B average by 0.25–0.83 points under identical rank, learning rate, and epoch budget (Appendix H.6). On the same ten datasets at 10^{-3}, final accuracy rises from 63.22% without polar reconstruction to 71.19% with it (Figure 3(a)). Averaging all shots of a class before merging is a different intervention. At 16 shots, individual-shot merging gains 1.9, 2.2, and 3.4 points across the three backbones (Figure 3(b)). It also makes K times as many updates per epoch. The result favors successive individual-gradient updates under equal example exposure (Appendix H.5).

6 Conclusion

OLM-SGD combines class-unique minibatching with orthogonal low-rank merging of individual-example class-specific gradients. It performs full fine-tuning and directly adapts all CLIP transformer blocks without additional trainable modules or adaptive moments. The analysis establishes retained-gradient alignment and energy control, with stationarity guarantees under explicit conditions. Across the evaluated few-shot settings, OLM-SGD achieves strong dataset-average accuracy and retains more of that accuracy during extended training. Matched trajectory controls support a contribution from gradient merging beyond the class uniqueness sampling technique or scalar norm control. Together, these results show that how support gradients are combined is consequential for both adaptation quality and sensitivity to the optimization configuration.

REFERENCES

Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. ProKeR: A kernel perspective on few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25092–25102, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Bendou_ProKeR_A_Kernel_Perspective_on_Few-Shot_Adaptation_of_Large_Vision-Language_CVPR_2025_paper.html.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 8694 of *Lecture Notes in Computer Science*, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29. URL https://link.springer.com/chapter/10.1007/978-3-319-10599-4_29.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014. URL https://openaccess.thecvf.com/content_cvpr_2014/html/Cimpoi_Describing_Textures_in_2014_CVPR_paper.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. CLIP's visual embedding projector is a few-shot cornucopia. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3254–3264, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Fahes_CLIPs_Visual_Embedding_Projector_is_a_Few-shot_Cornucopia_WACV_2026_paper.html.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007. doi: 10.1016/j.cviu.2005.09.012. URL https://doi.org/10.1016/j.cviu.2005.09.012.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132:581–595, 2024. doi: 10.1007/s11263-023-01891-x. URL https://link.springer.com/article/10.1007/s11263-023-01891-x.

Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18695–18705, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Gargiulo_Task_Singular_Vectors_Reducing_Task_Interference_in_Model_Merging_CVPR_2025_paper.html.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1225–1234. PMLR, 2016. URL https://proceedings.mlr.press/v48/hardt16.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected*

Topics in Applied Earth Observations and Remote Sensing, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079. URL https://doi.org/10.1137/0907079.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://doi.org/10.1080/01621459.1963.10500830.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2212.04089.

Keller Jordan. Muon: An optimizer for hidden layers in neural networks. Author blog, December 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77. URL https://doi.org/10.1109/ICCVW.2013.77.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9d27fdf2477ffbff837d73ef7ae23db9-Abstract.html.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

Colin McDiarmid. On the method of bounded differences. In J. Siemons (ed.), *Surveys in Combinatorics, 1989*, volume 141 of *London Mathematical Society Lecture Note Series*, pp. 148–188. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008. URL https://doi.org/10.1017/CBO9781107359949.008.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-3846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db36f4d603cc9e3a2a5e10b93e6428f2-Abstract-Conference.html.

Nairouz Mrabah, Nicolas Richet, Ismail Ben Ayed, and Eric Granger. Sparsity outperforms low-rank projections in few-shot adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3143–3152, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Mrabah_Sparsity_Outperforms_Low-Rank_Projections_in_Few-Shot_Adaptation_ICCV_2025_paper.html.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16428–16446. PMLR, 2022. URL https://proceedings.mlr.press/v162/navon22a.html.

Bac Nguyen, Stefan Uhlich, Fabien Cardinaux, Lukas Mauch, Marzieh Edraki, and Aaron Courville. SAFT: Towards out-of-distribution generalization in fine-tuning. In *Proceedings of the European Conference on Computer Vision (ECCV 2024)*, volume 15127 of *Lecture Notes in Computer Science*, pp. 138–154. Springer, 2025. doi: 10.1007/978-3-031-72890-7_9. URL https://link.springer.com/chapter/10.1007/978-3-031-72890-7_9.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20083–20093, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Senushkin_Independent_Component_Alignment_for_Multi-Task_Learning_CVPR_2023_paper.html.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. CLIPood: Generalizing CLIP to out-of-distributions. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31716–31731. PMLR, 2023. URL https://proceedings.mlr.press/v202/shu23a.html.

Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23681–23690, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Silva-Rodriguez_A_Closer_Look_at_the_Few-Shot_Adaptation_of_Large_Vision-Language_CVPR_2024_paper.html.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human action classes from videos in the wild. Technical Report CRCV-TR-12-01, Center for Research in Computer Vision, University of Central Florida, 2012. URL https://www.crcv.ucf.edu/research/data-sets/ucf101/.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How Muon's spectral design benefits generalization: A study on imbalanced data. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://iclr.cc/virtual/2026/poster/10008850.

Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7c40c5050bd029a3ea7ff8b01412f735-Abstract-Conference.html.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022a. URL https://proceedings.mlr.press/v162/wortsman22a.html.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010. doi: 10.1109/CVPR.2010.5539970. URL https://doi.org/10.1109/CVPR.2010.5539970.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. doi: 10.52202/075280-0310. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1644c9af28ab7916874f6fd6228a9bcf-Abstract-Conference.html.

Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6757–6767, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Yao_Visual-Language_Prompt_Tuning_With_Knowledge-Guided_Context_Optimization_CVPR_2023_paper.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 13695 of *Lecture Notes in Computer Science*, pp. 493–510. Springer, 2022. doi: 10.1007/978-3-031-19833-5_29. URL https://link.springer.com/chapter/10.1007/978-3-031-19833-5_29.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337–2348, 2022. doi: 10.1007/s11263-022-01653-1. URL https://link.springer.com/article/10.1007/s11263-022-01653-1.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15659–15669, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhu_Prompt-aligned_Gradient_for_Prompt_Tuning_ICCV_2023_paper.html.

Beier Zhu, Jiequan Cui, and Hanwang Zhang. Robust fine-tuning of zero-shot models via variance reduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 76967–76990, 2024. doi: 10.52202/079017-2448. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8cc7e1509fbfee9cabaacd3ab0bfe2b1-Abstract-Conference.html.

TABLE OF SUPPLEMENTARY CONTENTS

Appendix	Title	Page
A	**Few-shot Adaptation of CLIP**	**25**
B	**Extended Related Work**	**25**
C	**Algorithms**	**29**
C.1	Complete OLM-SGD Procedure	29
C.2	Control Optimizers	32
D	**Hyperparameters**	**39**
E	**Proofs and Additional Theoretical Results**	**45**
E.1	Setup, probability spaces, and empirical assumptions	45
E.2	OLM-specific geometry	46
E.3	Exact Energy Accounting and Factor Geometry	47
E.4	Quantitative Empirical Sufficient Conditions	48
E.5	Polar distortion under small factor overlap	49
E.6	Convergence from Quantified Retained-Gradient Errors	50
E.7	Conditional empirical convergence	51
E.8	A Polyak–Łojasiewicz Specialization	51
E.9	Exact Fixed-Parameter Sampling Variances	52
E.10	An Epoch-Wise Certificate for the Implemented Reshuffling	53
E.11	Quantified convergence for class-unique epochs	53
E.12	Population setup and trajectory-wise guarantees	56
E.13	A Direction-Preserving Population Certificate	57
E.14	Finite-Shot Gradient Error and Adaptive Iterates	57
E.15	A Population-Risk Generalization Bound	58
E.16	Independent Reference-Gradient Certification	59
E.17	Stability, Displacement, and Epoch Transfer	60
E.18	Numerical Reconstruction Errors	61
E.19	A Smooth Strongly Convex Counterexample	62
E.20	No Distribution-Free Exact Population Guarantee at Fixed Shot	63
F	**Comparison against State-of-the-Art PEFT Methods**	**63**
F.1	Few-Shot Accuracy across Backbones and Shot Levels	63
F.2	Cross-Dataset Transfer of the Selected Training Epoch	68
F.3	Long-Horizon Dynamics and Learning-Rate Sensitivity	70
F.4	Best-to-Final Test-Accuracy Drop	71
G	**Comparison against Robust Fine-Tuning Methods**	**78**
H	**Causal Trajectory Ablations**	**81**
H.1	Comparison against Adam and SGD	82
H.2	Comparison against SGD With Our Class-Unique Mini-Batching	95
H.3	Comparison against CTF-SGD	106
H.4	Comparison against GaLore	116
H.5	Comparison against OLM-SGD with Class-Mean Gradients	127
H.6	Comparison against OLM-SGD without Polar Reconstruction	130
I	**Rank Sensitivity**	**140**

SUMMARY OF NOTATION

The table records the notation used in the main paper and this supplementary material. Dimensions are symbolic when they depend on the CLIP architecture. Index symbols and proof-local quantities use their explicitly stated ranges. When one tensor is fixed, its superscript W is suppressed.

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
Indices, data, and objectives			
c, c'	Class indices in $\{1, \dots, C\}$	Class label and secondary class index in probability sums.	Sec. 3.1; Appendix A
i	Support index in $\{1, \dots, N\}$	Index of a labeled support example. Dataset indices in transfer figures are local to their captions.	Sec. 3.1
b, b'	Batch positions in $\{1, \dots, B_t\}$	Positions of individual gradient contributions in an update.	Eq. (23)
k	Integer in $\{1, \dots, K\}$	Round index in the class-unique iterator; shot index in class–shot gradient notation.	Eq. (22); Appendix E.9
j	Integer in $\{1, \dots, \lceil C/B \rceil\}$	Class-permutation chunk index. Dataset indices in transfer figures use their local ranges.	Eq. (22)
a	Integer in $\{1, \dots, r_W\}$	Retained singular-component index for tensor W.	Eq. (25)
t, t'	Nonnegative iteration indices	Current update index and auxiliary iteration index in trajectory sums.	Eq. (11); Appendix E.10
e	Epoch index	Training pass index, usually $1 \le e \le E$; evaluation grids may include $e = 0$.	Algorithm 2; Appendix E.17
C	Integer, $C \ge 2$	Number of classes in the adaptation task.	Sec. 3.1
K	Integer, $K \ge 1$	Number of labeled support examples per class.	Sec. 3.1
$N = CK$	Positive integer	Total number of support examples; not the number of classes.	Sec. 3.1
$\mathcal{D}_{\mathrm{sup}}$	Set of N labeled examples	Balanced support set $\{(x_i, y_i)\}_{i=1}^{N}$ reused throughout adaptation.	Sec. 3.1
x_i	Input image	Image component of support example i; x denotes a generic image.	Sec. 3.1
y_i	Integer in $\{1, \dots, C\}$	Class label of image x_i.	Sec. 3.1
$\ell_i(\Theta)$ $\ell(\Theta; x, c)$	Scalar-valued losses	Full C-way cross-entropy: $-\log p_\Theta(y_i \mid x_i)$ for a support example and $-\log p_\Theta(c \mid x)$ for a generic labeled image.	Eq. (21); Appendix E.12
F	$\mathbb{R}^d \to \mathbb{R}$	Empirical support objective $N^{-1} \sum_i \ell_i(\Theta)$.	Eq. (1)
$\mathcal{L}(\Theta_t)$	Real scalar	Mean loss over the current B_t examples; distinct from the complete-support objective F.	Eq. (2)
$\mathcal{P}_c$	Image probability distribution	Population image distribution conditioned on class c.	Appendix E.12
$\mathcal{R}$	$\mathbb{R}^d \to \mathbb{R}$	Class-balanced population risk $C^{-1} \sum_c \mathbb{E}_{x \sim \mathcal{P}_c}[\ell(\Theta; x, c)]$.	Eq. (105)
CLIP representations and class probabilities			
$f_{\mathrm{img}}(x; \Theta)$	Unit-norm embedding vector	Normalized CLIP image-encoder output in the shared image–text embedding space.	Appendix A
$f_{\mathrm{txt}}(x^{\mathrm{txt}}; \Theta)$	Unit-norm embedding vector	Normalized CLIP text-encoder output in the same shared embedding space.	Appendix A
x^{txt}	Text input	Input prompt or text sequence supplied to the text encoder.	Appendix A

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
$\mathcal{X}_c^{\mathrm{prompt}}$	Finite, nonempty set of text prompts	Class-name prompts used to construct the text prototype of class c.	Appendix A
$\bar{z}_c^{\mathrm{txt}}(\Theta)$ $z_c^{\mathrm{txt}}(\Theta)$	Shared-dimension embedding vectors	Mean prompt embedding and its unit-norm normalization, respectively. Normalization requires a nonzero mean.	Eq. (19)
τ	Positive real scalar	Softmax temperature for normalized image–text similarities.	Eq. (20)
$p_\Theta(c \mid x)$	Scalar in $[0, 1]$	Predicted probability of class c. The C class probabilities sum to one.	Eq. (20)
Parameters, minibatches, and training budgets			
$\Theta, \Theta_t, \Theta_0$	Vectors in $\mathbb{R}^d$	Trainable parameter vector, its value at iteration t, and the pretrained initialization.	Sec. 3.1
d	Positive integer	Number of trainable scalar coordinates: $d = \sum_{W \in \mathcal{W}} m_W n_W$.	Appendix C.1
$\mathcal{W}$ W, W_t	Ordered tensor collection; a tensor and its current value	Fixed collection of trainable parameter tensors. W identifies one tensor and W_t is its value at iteration t.	Sec. 3.1
m_W, n_W	Positive integers	Row and column dimensions of the fixed matrix representation of tensor W.	Eq. (4)
mat_W mat_W^{-1}	Entry-preserving reshape maps	Map between the native tensor shape and $\mathbb{R}^{m_W \times n_W}$. Vectors become columns and scalars become 1×1 matrices.	Eq. (4)
B	Positive integer	Minibatch capacity in examples. Class-unique OLM-SGD requires $B \leq C$; randomized batching has no such class-count restriction.	Sec. 3.1; Appendix C
$\mathcal{B}_t$ $B_t = \lvert\mathcal{B}_t\rvert$	Minibatch; integer $1 \leq B_t \leq B$	Realized example group and its actual size. Incomplete batches use B_t, not the capacity B, for averaging.	Sec. 3.1
$i_{t,b}$	Index in $\{1, \ldots, N\}$	Support-set index of the example at position b in update t.	Sec. 3.1
$\mathcal{I}_c$	Ordered list of K support indices	Indices of examples with label c, shuffled independently at each epoch.	Sec. 3.1
π_k	Permutation of $\{1, \ldots, C\}$	Fresh class permutation in round k of an epoch.	Eq. (22)
$\mathcal{C}_{k,j}$ $\mathcal{B}_{k,j}$	Class subset; associated example group	Classes in chunk j of round k and one selected support example per class, processed in permutation order.	Eq. (22)
E T	Positive integers	Number of complete support-set epochs and number of parameter updates. For OLM-SGD, $T = EK\lceil C/B \rceil$.	Algorithm 2; Appendix E.10
η_t η	Positive real scalars	Learning rate at update t and its constant-step specialization $\eta_t = \eta$.	Eq. (11); Eq. (80)
Λ_T	Positive real scalar	Cumulative learning rate $\sum_{t=0}^{T-1} \eta_t$.	Sec. 4
Individual gradients and orthogonal low-rank merging			
$g_t^{(b)}$	Vector in $\mathbb{R}^d$	Individual-example gradient $\nabla_\Theta \ell_{i_{t,b}}(\Theta_t)$. Class uniqueness makes its label distinct within the batch.	Eq. (2)
g_t	Vector in $\mathbb{R}^d$	Ordinary minibatch mean gradient $B_t^{-1} \sum_b g_t^{(b)}$, used by SGD.	Eq. (2)
$g_t^{(b),W}$ g_t^W	Native-shaped tensor gradients	Slices of the individual gradient and the minibatch mean associated with tensor W.	Eq. (4); Eq. (31)
$G_t^{(b),W}$	Matrix in $\mathbb{R}^{m_W \times n_W}$	Matrix representation $\mathrm{mat}_W(g_t^{(b),W})$ of an individual tensor gradient.	Eq. (4)

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
r r_W	Positive integers	Global gradient-rank budget and effective tensor rank $r_W = \min\{r, m_W, n_W\}$. Neither is a weight-parameterization rank.	Eq. (24)
Π_{r_W}	Matrix-valued operator	Exact truncated SVD retaining r_W components. Numerical approximations are distinguished in the appendix.	Eq. (5); Appendix E.2
$U_t^{(b)}$ $V_t^{(b)}$	$m_W \times r_W$ and $n_W \times r_W$ matrices	Retained left and right singular-vector factors for a fixed tensor; each has orthonormal columns.	Eq. (5)
$\Sigma_t^{(b)}$	Diagonal $r_W \times r_W$ matrix	Retained nonnegative singular values, ordered from largest to smallest.	Eq. (5)
$\sigma_{t,a}^{(b)}$	Nonnegative real scalar	Retained singular value of component a for batch contribution b.	Eq. (25)
$u_{t,a}^{(b)}$ $v_{t,a}^{(b)}$	Vectors in $\mathbb{R}^{m_W}$ and $\mathbb{R}^{n_W}$	Corresponding columns of the retained left and right factors.	Eq. (25)
U_t V_t	$m_W \times R_t$ and $n_W \times R_t$ matrices	Concatenated retained factors in the same batch order. Cross-class columns need not be orthogonal.	Eq. (6)
R_t R	Positive integers	Concatenated width $R_t = r_W B_t$ for one tensor. Proofs use R when iteration and tensor indices are suppressed.	Eq. (6); Appendix E.2
s_t	Vector in $\mathbb{R}^{R_t}_{\geq 0}$	Concatenated retained singular values divided once by the actual batch size B_t.	Eq. (7)
$\widetilde{G}_t$	Matrix in $\mathbb{R}^{m_W \times n_W}$	Mean of the individually truncated matrices before polar reconstruction: $U_t \operatorname{diag}(s_t) V_t^\top$.	Eq. (8)
A m_A, n_A, q_A	Generic matrix; positive integer dimensions	Input to the rectangular polar operation, with $A \in \mathbb{R}^{m_A \times n_A}$ and $q_A = \min(m_A, n_A)$.	Eq. (9)
P_A, D_A, Q_A	Reduced-SVD factors	$A = P_A D_A Q_A^\top$, with P_A of size $m_A \times q_A$, Q_A of size $n_A \times q_A$, and diagonal D_A of size $q_A \times q_A$.	Eq. (9)
$\operatorname{polar}(A)$	Matrix in $\mathbb{R}^{m_A \times n_A}$	SVD-selected factor $P_A Q_A^\top$. All reduced-SVD columns are retained, including zero-singular-value directions.	Eq. (9)
$U_{t,\mathrm{pol}}$ $V_{t,\mathrm{pol}}$	Same shapes as U_t, V_t	Polar factors of the concatenations. Each has orthonormal columns when tall and orthonormal rows when wide.	Eq. (26)
$\widehat{G}_t$	Matrix in $\mathbb{R}^{m_W \times n_W}$	OLM merged tensor matrix $U_{t,\mathrm{pol}} \operatorname{diag}(s_t) V_{t,\mathrm{pol}}^\top$.	Eq. (10)
$\widetilde{g}_t$	Vector in $\mathbb{R}^d$	Concatenation of restored retained means over all tensors; update direction of the without-polar ablation.	Appendix E.4; Eq. (46)
$\widehat{g}_t$	Vector in $\mathbb{R}^d$	Concatenation of restored merged tensors used in $\Theta_{t+1} = \Theta_t - \eta_t \widehat{g}_t$.	Eq. (11)
$\mathcal{Q}_t$	Nonnegative real scalar	Retained-energy budget $\sum_W \|s_t^W\|_2^2$. It bounds $\|\widehat{g}_t\|_2^2$ but is not always equal to it.	Eq. (49)
Optimizer states and ablation-specific quantities			
$\mathsf{Mom}_{t,\mathrm{Adam}}^{(1)}$ $\mathsf{Mom}_{t,\mathrm{Adam}}^{(2)}$	Vectors in $\mathbb{R}^d$	Adam first- and second-moment buffers before update t, initialized to zero.	Algorithm 3
$\widehat{\mathsf{Mom}}_{t+1,\mathrm{Adam}}^{(1)}$ $\widehat{\mathsf{Mom}}_{t+1,\mathrm{Adam}}^{(2)}$	Vectors in $\mathbb{R}^d$	Bias-corrected Adam moments after incorporating the current gradient.	Algorithm 3
β_1, β_2	Scalars in $[0, 1)$	Exponential decay coefficients for first and second moments in Adam and GaLoRE.	Appendix C

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
$\varepsilon_{\mathrm{Adam}}$ $\varepsilon_{\mathrm{GaL}}$	Positive real scalars	Denominator stabilization constants in Adam and GaLoRE, respectively.	Algorithm 3; Eq. (32)
r_{GaL} r_W^{GaL}	Positive integers	GaLoRE projection-rank budget and $\min\{r_{\mathrm{GaL}}, m_W, n_W\}$. Applied after gradient averaging.	Eq. (30)
$Q_{t,\mathrm{GaL}}^W$	One-sided projection basis	Right singular vectors of size $n_W \times r_W^{\mathrm{GaL}}$ if $m_W \geq n_W$; left vectors of size $m_W \times r_W^{\mathrm{GaL}}$ otherwise.	Eq. (31)
T_{proj}	Positive integer	Number of updates between GaLoRE projection-basis refreshes; the basis is also initialized at $t = 0$.	Appendix C; Table 15
$G_{t,\mathrm{proj}}^W$	Projected gradient matrix	Size $m_W \times r_W^{\mathrm{GaL}}$ if $m_W \geq n_W$, and $r_W^{\mathrm{GaL}} \times n_W$ otherwise.	Eq. (31)
$\mathsf{Mom}_{t,\mathrm{GaL}}^{(1),W}$ $\mathsf{Mom}_{t,\mathrm{GaL}}^{(2),W}$	Same shape as $G_{t,\mathrm{proj}}^W$	GaLoRE first- and second-moment buffers in projected coordinates, retained across basis refreshes.	Appendix C; Eq. (32)
D_t^W	Same shape as $G_{t,\mathrm{proj}}^W$	Adam-preconditioned GaLoRE direction in projected coordinates.	Eq. (32)
Δ_t^W	Matrix in $\mathbb{R}^{m_W \times n_W}$	GaLoRE direction lifted to the original matrix space, before multiplication by the learning rate and subtraction.	Eq. (33)
$\mathcal{E}_t^{\mathrm{CTF}}$	Nonnegative real scalar	Full-gradient self-energy $B_t^{-2} \sum_b \|g_t^{(b)}\|_2^2$, without pairwise cross terms.	Eq. (38)
$\mathcal{B}_{t,\mathrm{L}}$ $\mathcal{B}_{t,\mathrm{R}}$	Nonempty microbatches for $B_t \geq 2$	Uniform random partition of the realized minibatch used to estimate its self-energy.	Appendix C; Eq. (40)
$B_{t,\mathrm{L}}$ $B_{t,\mathrm{R}}$	Positive integer microbatch sizes	$B_{t,\mathrm{L}} = \lfloor B_t/2 \rfloor$ and $B_{t,\mathrm{R}} = B_t - B_{t,\mathrm{L}}$.	Appendix C
$g_{t,\mathrm{L}}$ $g_{t,\mathrm{R}}$	Vectors in $\mathbb{R}^d$	Microbatch mean gradients evaluated at the same parameters. Their size-weighted mean recovers g_t.	Eq. (40)
$\widehat{\mathcal{E}}_t^{\mathrm{CTF}}$	Nonnegative real scalar	Two-microbatch estimator of the full-gradient self-energy; unbiased over uniform splits for fixed individual gradients.	Eq. (41)
α_t^{CTF}	Scalar in $[0, 1]$	Global shrink coefficient multiplying g_t in CTF-SGD; it does not change the SGD direction.	Eq. (43)
$\varepsilon_{\mathrm{CTF}}$	Positive real scalar	Numerical tolerance in the CTF-SGD coefficient and zero-norm check.	Eq. (43)
λ	Nonnegative real scalar	Weight-decay coefficient in the Adam/SGD hyperparameter table; set to zero in these comparisons.	Table 14
Class-mean gradient ablation			
B_{cls}	Integer, $1 \leq B_{\mathrm{cls}} \leq C$	Update capacity in complete classes, not individual images.	Appendix C
$\mathcal{C}_t^{\mathrm{cm}}$ M_t^{cm}	Ordered class group; its integer size	Selected classes and their count $M_t^{\mathrm{cm}} = \|\mathcal{C}_t^{\mathrm{cm}}\| \leq B_{\mathrm{cls}}$.	Appendix C
$\mathcal{B}_t^{\mathrm{cm}}$	Group of support indices	All support examples from the selected classes: $\bigcup_{c \in \mathcal{C}_t^{\mathrm{cm}}} \mathcal{I}_c$, containing KM_t^{cm} examples.	Appendix C
$\mu_c^W(\Theta_t)$	Native-shaped tensor gradient	Slice of the complete K-shot class mean $\mu_c(\Theta_t)$ associated with W.	Eq. (47); Eq. (83)

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
$U^{(c),W}_{t,\mathrm{cm}}$ $\Sigma^{(c),W}_{t,\mathrm{cm}}$ $V^{(c),W}_{t,\mathrm{cm}}$	Rank-r_W SVD factors	Retained factors of the class-mean matrix, with respective sizes $m_W \times r_W$, $r_W \times r_W$, and $n_W \times r_W$.	Eq. (47)
$U^{W}_{t,\mathrm{cm}}$ $V^{W}_{t,\mathrm{cm}}$	Concatenated factor matrices	Factors concatenated in selected-class order, with width $r_W M^{\mathrm{cm}}_t$ and m_W or n_W rows.	Eq. (48)
$s^{W}_{t,\mathrm{cm}}$	Nonnegative vector of length $r_W M^{\mathrm{cm}}_t$	Concatenated class-mean singular values divided by M^{cm}_t. The class gradient already includes $1/K$.	Eq. (48)
$\widehat{G}^{W}_{t,\mathrm{cm}}$	Matrix in $\mathbb{R}^{m_W \times n_W}$	Polar reconstruction from the retained class-mean factors and their scaled singular values.	Eq. (48)
Conditional convergence, update bias, and distortion			
$\mathcal{H}_t$	Conditioning history	Fixed support and all training randomness revealed before update t; epoch-start history precedes that epoch's fresh shuffles.	Appendix E.1
$\mathbb{E}_t[\cdot]$	Conditional expectation	$\mathbb{E}[\cdot \mid \mathcal{H}_t]$. Unless stated otherwise, expectations over training hold the support fixed.	Appendix E.1
L_F	Positive real scalar	Lipschitz constant of the empirical gradient on the relevant update region.	Assumption E.1
$L_{\mathcal{R}}$	Nonnegative real scalar	Lipschitz constant of the population gradient on the relevant update segments.	Theorem E.17
$F_{\inf}$ $\mathcal{R}_{\inf}$	Finite real lower bounds	Lower bounds for the empirical objective and population risk, respectively.	Assumption E.1; Theorem E.17
Δ_F $\Delta_{\mathcal{R}}$	Nonnegative real scalars	Initial gaps $F(\Theta_0) - F_{\inf}$ and $\mathcal{R}(\Theta_0) - \mathcal{R}_{\inf}$.	Theorems E.10 and E.17
α_{olm} β_{olm}	Scalars $\alpha_{\mathrm{olm}} > 0$, $\beta_{\mathrm{olm}} \geq 0$	Multiplicative alignment coefficient and additive alignment deficit in the conditional empirical assumption.	Eq. (50)
M_{olm} σ_{olm}	Nonnegative real scalars	Gradient-dependent moment coefficient and additive moment scale; the additive term is σ^2_{olm}.	Eq. (51)
μ_{PL}	Positive real scalar	Polyak–Łojasiewicz constant in the optional empirical objective-gap analysis.	Eq. (81)
$e_{\sup}(\Theta)$	Vector in $\mathbb{R}^d$	Finite-support gradient discrepancy $\nabla F(\Theta) - \nabla \mathcal{R}(\Theta)$. Fixed as a function of Θ once the support is drawn.	Eq. (106)
ξ_t	Vector in $\mathbb{R}^d$	Batch error $g_t - \nabla F(\Theta_t)$; not assumed conditionally centered under reshuffling.	Eq. (106)
$\mathfrak{b}_t$	Nonnegative real scalar	Conditional empirical update bias $\|\mathbb{E}_t[\widehat{g}_t] - \nabla F(\Theta_t)\|_2$.	Eq. (60)
$\varepsilon_{t,\mathrm{tail}}$	Nonnegative real scalar	Tensor-wise aggregate of mean discarded-gradient norms; upper-bounds $\|\widetilde{g}_t - g_t\|_2$.	Eq. (61)
$\varepsilon_{t,\mathrm{polar}}$	Nonnegative real scalar	Polar reconstruction distortion $\|\widehat{g}_t - \widetilde{g}_t\|_2$, computed from tensor Frobenius differences.	Eq. (62)
$\varepsilon_{t,\mathrm{sample}}$	Nonnegative real scalar	Conditional sampling bias $\|\mathbb{E}_t[g_t] - \nabla F(\Theta_t)\|_2$; distinct from sampling variance.	Eq. (63)
ρ_{olm} δ_{olm}	$0 \leq \rho_{\mathrm{olm}} < 1$, $\delta_{\mathrm{olm}} \geq 0$	Relative and additive upper bounds on the combined conditional merge and sampling bias.	Eq. (65)

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
$B_{\min}$	Integer, $1 \le B_{\min} \le B_t$	Lower bound on realized minibatch sizes in the component-growth argument.	Lemma E.6
A_{grad} ν_{grad}	Nonnegative real scalars	Component-gradient moment constants, giving $M_{\mathrm{olm}} = A_{\mathrm{grad}}/B_{\min}$ and $\sigma^2_{\mathrm{olm}} = \nu^2_{\mathrm{grad}}/B_{\min}$.	Eqs. (68) and (69)
$\varepsilon_{t,\mathrm{target}}$	Nonnegative real scalar	Retained-target error $\|\widetilde{g}_t - \nabla F(\Theta_t)\|_2$; a certified upper bound can replace it in the inequalities.	Eq. (70); Theorem E.8
ρ_{ret} κ_{ret}	$0 \le \rho_{\mathrm{ret}} < 1$, $\kappa_{\mathrm{ret}} \ge 0$	Relative retained-target error and retained-energy scales, compared with $\|\nabla F(\Theta_t)\|_2$, under the stated pathwise or conditional RMS conditions.	Eq. (71); Proposition E.9
f_t	Vector in $\mathbb{R}^d$	Full empirical gradient $\nabla F(\Theta_t)$ in the direct geometric analysis.	Theorem E.8
δ_{ret} ν_{ret}	Nonnegative real scalars	Additive conditional RMS bounds for retained-target error and retained energy.	Eq. (74)
a_{ret} c_{ret}	Real scalar; nonnegative scalar	Quadratic alignment coefficient and linear error coefficient derived from the RMS conditions; a useful certificate requires $a_{\mathrm{ret}} > 0$.	Proposition E.9
Class-wise sampling and epoch reshuffling			
$g_{c,k}(\Theta)$	Vector in $\mathbb{R}^d$	Support-example gradient reindexed by class c and shot k.	Appendix E.9
$\mu_c(\Theta)$ $\mu(\Theta)$	Vectors in $\mathbb{R}^d$	Empirical class mean $K^{-1}\sum_k g_{c,k}(\Theta)$ and support mean $C^{-1}\sum_c \mu_c(\Theta) = \nabla F(\Theta)$.	Eq. (83)
$\Gamma_{\mathrm{within}}(\Theta)$ $\Gamma_{\mathrm{between}}(\Theta)$	Nonnegative real scalars	Mean squared within-class support-gradient dispersion and dispersion of empirical class means around the full support mean.	Eq. (85)
$g_{\mathrm{class}}(\Theta)$	Random vector in $\mathbb{R}^d$	Ordinary gradient mean for a fresh draw of distinct classes and one uniformly sampled support example per class at fixed Θ.	Proposition E.12
$\mu_c^{\mathrm{pop}}(\Theta)$ $\mu^{\mathrm{pop}}(\Theta)$	Vectors in $\mathbb{R}^d$	Population class-gradient mean and their class-balanced mean $\nabla\mathcal{R}(\Theta)$.	Eq. (113); Appendix E.9
$\Gamma^{\mathrm{pop}}_{\mathrm{within}}(\Theta)$	Nonnegative real scalar	Class-averaged population gradient variance within classes.	Eq. (113)
$\Gamma^{\mathrm{pop}}_{\mathrm{between}}(\Theta)$	Nonnegative real scalar	Dispersion of population class-gradient means around their class-balanced mean.	Eq. (90)
J	Positive integer	OLM-SGD updates per epoch: $J = K\lceil C/B \rceil$.	Appendix E.10
t_e $\mathcal{T}_e$	Iteration index; set of J indices	Start $t_e = (e-1)J$ and update indices $\{t_e, \dots, t_e + J - 1\}$ of epoch e.	Appendix E.10
ω_e	Vector in $\mathbb{R}^d$	Frozen-parameter remainder from equally weighting batch means rather than examples; zero when B divides C.	Eq. (91)
q_{batch} b_j	Positive integers	Number $\lceil C/B \rceil$ of chunks per round and realized chunk size $\min\{B, C-(j-1)B\}$.	Appendix E.11
a_{batch}	Nonnegative real scalar	Incomplete-batch weighting factor; zero for equal-sized chunks.	Eq. (93)
$\mathbb{E}_e[\cdot]$	Conditional expectation	Conditioning on $\mathcal{H}_{t_e}$ before the epoch shuffles, with the support fixed.	Proposition E.14
η_e γ_e	Positive real scalars	Per-update step constant within epoch e, and cumulative epoch step $\gamma_e = J\eta_e$.	Appendix E.11
$\Lambda_E^{\mathrm{epoch}}$	Positive real scalar	Sum $\sum_{e=1}^E \gamma_e$ of epoch steps.	Theorem 4.2

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
f_e h_e	Vectors in $\mathbb{R}^d$	Full empirical gradient at epoch start and arithmetic mean of the epoch's realized OLM updates.	Appendix E.11
u_e D_e, d_e	Vector in $\mathbb{R}^d$; nonnegative scalars	Residual $h_e - f_e - \omega_e$, its norm, and its conditional RMS norm.	Eq. (98)
ν_e	Nonnegative real scalar	Conditional RMS norm of the frozen incomplete-batch remainder ω_e.	Eq. (94)
M_e c_{drift}	Nonnegative real scalars	Epoch-mean norm of the merge change and coefficient bounding within-epoch drift by $c_{\mathrm{drift}}\gamma_e$.	Eq. (103)
m, v	Nonnegative real scalars	Uniform RMS bounds on M_e and ν_e in the epoch-rate corollary.	Corollary E.15
ρ_{epoch} δ_e	$0 \le \rho_{\mathrm{epoch}} < 1$; nonnegative scalar	Relative and additive conditional RMS residual bounds; δ_e is epoch-history measurable.	Corollary E.16
$\gamma_{\max}$ c_{epoch}	Positive real scalars	Epoch-step upper bound and remaining descent coefficient $1 - (1 + L_F\gamma_{\max})\rho_{\mathrm{epoch}}^2$.	Corollary E.16
$\bar{\delta}$	Nonnegative real scalar	Uniform RMS bound on the additive epoch residual δ_e.	Corollary E.16
Finite-shot concentration and independent reference certification			
$\mathcal{A}$	Compact subset of $\mathbb{R}^d$	Deterministic parameter region independent of the support draw, used for uniform gradient and risk bounds.	Proposition E.19
R_{loc}	Positive real scalar	Radius of a support-independent ball around Θ_0; distinct from the concatenated matrix width R.	Eq. (118)
H_{grad}	Nonnegative real scalar	Uniform bound on per-example gradient norms in the stated region; also a loss-Lipschitz bound when assumed.	Proposition E.19; Proposition E.20
L_ℓ	Nonnegative real scalar	Lipschitz constant of per-example gradients on the relevant region or segments.	Proposition E.19; Proposition E.13
$\varepsilon_{\mathrm{cov}}$	Positive real scalar	Euclidean covering radius for the deterministic parameter region.	Proposition E.19
$\mathcal{N}_{\mathrm{cov}}$	Positive integer	Number of centers in the deterministic $\varepsilon_{\mathrm{cov}}$-cover of $\mathcal{A}$.	Proposition E.19
δ_{conf}	Scalar in $(0, 1)$	Failure-probability level in the stated concentration or reference certificate.	Proposition E.19; Proposition E.21
$\varepsilon_{\mathrm{shot}}$	Nonnegative real scalar	Uniform finite-support gradient-error radius combining statistical error and covering error.	Eq. (116)
u_{dev}	Positive real scalar	Local deviation threshold in gradient and risk concentration inequalities.	Eq. (117); Appendix E.15
M_{loss}	Positive real scalar	Uniform loss upper bound, with losses in $[0, M_{\mathrm{loss}}]$, for the separate risk-generalization result.	Proposition E.20
$\varepsilon_{\mathrm{risk}}$	Nonnegative real scalar	Uniform bound on $\lvert F(\Theta) - \mathcal{R}(\Theta)\rvert$ over $\mathcal{A}$ on the stated risk event.	Eq. (121)
$\mathcal{D}_{\mathrm{ref}}$	Independent labeled audit set	Reference examples independent of support data and all trajectory-construction and model-selection decisions.	Appendix E.16
K_{ref} N_{ref}	Positive integers	Reference examples per class and total reference size $N_{\mathrm{ref}} = CK_{\mathrm{ref}}$.	Appendix E.16

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
F_{ref}	$\mathbb{R}^d \to \mathbb{R}$	Empirical reference objective with the same loss and class weights as F; not used for training.	Appendix E.16
$\varepsilon_{\mathrm{ref},T}$	Nonnegative real scalar	Simultaneous reference-gradient concentration radius for T audited parameter points.	Eq. (124)
Polar geometry, stability, and proof-local constructions			
S_t^W S	Diagonal $R_t \times R_t$ matrix; scoped version	Coefficient matrix $\mathrm{diag}(s_t^W)$; iteration and tensor indices are suppressed when fixed.	Eq. (59)
$\mathsf{P}_{U,t}^W$ $\mathsf{P}_{V,t}^W$	Orthogonal $R_t \times R_t$ projectors	Gram matrices of the completed polar factors; scoped versions are P_U and P_V.	Proposition E.4
U_{num} V_{num}	Matrices of the retained-factor shapes	Numerically perturbed polar factors, with iteration and tensor indices suppressed.	Proposition E.22
G_{num} $g_{t,\mathrm{num}}$	$m_W \times n_W$ matrix; vector in $\mathbb{R}^d$	Numerical tensor reconstruction and assembled numerical update.	Proposition E.22
a_U, a_V	Nonnegative real scalars	Spectral-norm bounds on the numerical polar-factor errors.	Proposition E.22
z_W z_t	Nonnegative real scalars	Per-tensor numerical reconstruction error bound and its aggregate Euclidean bound.	Eq. (139)
H_U, H_V	Positive-semidefinite $R \times R$ matrices	Gram square roots $(U^\top U)^{1/2}$ and $(V^\top V)^{1/2}$ used in the polar-alignment proof.	Eq. (52)
χ_U, χ_V	Scalars in $[0,1)$ in the overlap bound	Spectral deviations $\|U^\top U - I_R\|_2$ and $\|V^\top V - I_R\|_2$ for tall, full-column-rank factors.	Appendix E.5
$\psi(\chi)$	$[0,1) \to \mathbb{R}_{\geq 0}$	Sharpened polar perturbation bound $1 - \sqrt{1-\chi}$.	Appendix E.5
κ	Scalar in $(0,1)$; also examined at zero	Local parameter of the strongly convex counterexample, not the retained-energy constant κ_{ret}.	Appendix E.19
$\overline{G}$	2×2 matrix in the counterexample	Ordinary mean of the three explicitly constructed example-gradient matrices.	Appendix E.19
ϕ	Scalar-valued image feature	Fixed feature with values $0, +1, -1$ in the two-class impossibility construction; not a CLIP embedding.	Appendix E.20
$\mathcal{R}^+, \mathcal{R}^-$	$\mathbb{R} \to \mathbb{R}$	Population risks of the paired scalar cross-entropy problems with opposite nonzero feature signs.	Eq. (147)
ϑ	Angle in $(0, \pi/2)$	Local parameter in the example showing lack of uniform continuity near a rank-deficient concatenation.	Eq. (126)
t_0, t_1	Integers, $0 \leq t_0 < t_1 \leq T$	Endpoints of a trajectory interval used for parameter, loss, and accuracy drift bounds.	Eq. (132)
L_{prob}	Positive real scalar	Uniform parameter-Lipschitz constant for each class probability in the accuracy-drift argument.	Appendix E.17
$\mathsf{m}_\Theta(x, c)$	Scalar in $[-1,1]$	True-class probability margin $p_\Theta(c \mid x) - \max_{c' \neq c} p_\Theta(c' \mid x)$.	Appendix E.17
$\mathrm{Err}(\Theta)$	Scalar in $[0,1]$	Class-balanced population misclassification probability.	Eq. (135)
$\mathcal{E}_{\mathrm{eval}}$	Finite subset of $\{0, \ldots, E\}$ containing zero	Prespecified epoch grid for the conditional epoch-transfer certificate.	Appendix E.17
$\mathsf{Risk}_{\mathrm{src}}(e)$ $\mathsf{Risk}_{\mathrm{tgt}}(e)$	Scalar-valued epoch profiles	Population evaluation risks along separately trained source- and target-task trajectories.	Eqs. (136) and (137)

Continued on the next page

Symbol	Type / Domain / Dimension	Definition or Role	Canonical Definition Location
$\widehat{\mathsf{Risk}}_{\mathrm{src}}(e)$	Scalar validation estimate	Independent-validation estimate of the source-task risk at epoch e.	Eq. (136)
$\varepsilon_{\mathrm{val}}$ Δ_{curve}	Nonnegative real scalars	Uniform source-validation error and discrepancy between source and target risk profiles centered at epoch zero.	Eqs. (136) and (137)
$\widehat{e}$	Element of $\mathcal{E}_{\mathrm{eval}}$	Epoch minimizing estimated source-validation risk on the prescribed grid.	Eq. (138)
Norms, matrix operations, and ordering conventions			
$\|\cdot\|_2$	Nonnegative real-valued norm	Euclidean norm for vectors and spectral norm for matrices.	Appendix C.1
$\|\cdot\|_F$	Nonnegative real-valued matrix norm	Frobenius norm; entry-preserving tensor reshaping preserves its corresponding Euclidean energy.	Appendix C.1
$\langle\cdot,\cdot\rangle$ $\langle\cdot,\cdot\rangle_F$	Real-valued inner products	Euclidean inner product for vectors and the sum of entrywise products for equally shaped matrices.	Eq. (13); Eq. (54)
$\mathrm{diag}(\cdot)$	Diagonal-matrix construction	Forms a square diagonal matrix from the entries of its vector argument.	Eq. (7)
vec concat	Vectorization and concatenation	List tensor entries and join tensor vectors in the fixed parameter order of $\mathcal{W}$.	Eq. (27)
$\mathrm{tr}(\cdot)$	Real-valued matrix operator	Trace of a square matrix.	Eq. (54)
I_R $\preceq$	Identity matrix; matrix order	$R \times R$ identity and positive-semidefinite order. Identity subscripts specify their dimensions.	Appendices E.2 and E.5
$(u)_+$	Nonnegative real scalar	Positive part $\max(u, 0)$ of a real scalar u.	Eq. (70)

A FEW-SHOT ADAPTATION OF CLIP

Vision–language models learn a shared embedding space for visual and textual inputs. CLIP (Radford et al., 2021) uses an image encoder and a text encoder. We denote by $f_{\mathrm{img}}(x;\Theta)$ and $f_{\mathrm{txt}}(x^{\mathrm{txt}};\Theta)$ their ℓ_2-normalized outputs, where x^{txt} denotes a text input. Both maps use the trainable parameter vector Θ. CLIP aligns paired image–text embeddings during pretraining through a symmetric InfoNCE loss (van den Oord et al., 2018).

Zero-shot inference. For each of the C classes, let $\mathcal{X}_c^{\mathrm{prompt}}$ be a finite, nonempty set of class-name prompts. The mean text embedding and normalized class prototype are

$$\begin{aligned} \bar{z}_c^{\mathrm{txt}}(\Theta) &= \frac{1}{|\mathcal{X}_c^{\mathrm{prompt}}|} \sum_{x^{\mathrm{txt}} \in \mathcal{X}_c^{\mathrm{prompt}}} f_{\mathrm{txt}}(x^{\mathrm{txt}};\Theta), \\ z_c^{\mathrm{txt}}(\Theta) &= \frac{\bar{z}_c^{\mathrm{txt}}(\Theta)}{\|\bar{z}_c^{\mathrm{txt}}(\Theta)\|_2}. \end{aligned} \tag{19}$$

The normalization assumes a nonzero mean embedding. Averaging combines the available descriptions of a class without adding separate classifier parameters. With one prompt, the prototype is its normalized text embedding. For an image x and softmax temperature $\tau > 0$, the probability $p_\Theta(c \mid x)$ from Section 3.1 is

$$p_\Theta(c \mid x) = \frac{\exp\big(f_{\mathrm{img}}(x;\Theta)^\top z_c^{\mathrm{txt}}(\Theta)/\tau\big)}{\sum_{c'=1}^{C} \exp\big(f_{\mathrm{img}}(x;\Theta)^\top z_{c'}^{\mathrm{txt}}(\Theta)/\tau\big)}. \tag{20}$$

Since both embeddings have unit norm, their inner product is a cosine similarity. Thus, class scores compare embedding directions rather than raw feature magnitudes. Zero-shot inference evaluates this rule at Θ_0 without using labeled support images.

Few-shot adaptation. We use the balanced support set $\mathcal{D}_{\mathrm{sup}}$ from Sec. 3.1, with $N = CK$ labeled images and integer labels $y_i \in \{1, \dots, C\}$. The experiments consider $K \in \{1, 2, 4, 8, 16\}$. Starting from Θ_0, adaptation optimizes the existing objective F in (1):

$$\ell_i(\Theta) = -\log p_\Theta(y_i \mid x_i), \qquad F(\Theta) = \frac{1}{N} \sum_{i=1}^{N} \ell_i(\Theta). \tag{21}$$

This supervised classification loss differs from the image–text matching objective used during pretraining.

At iteration t, the minibatch loss $\mathcal{L}(\Theta_t)$ averages the losses of the B_t examples in $\mathcal{B}_t$, as defined in (2). Each loss still compares its image against all C class prototypes. Thus, class-unique batching changes the gradient contributions to an update, not the classification task. When $K > 1$, different shots of a class enter different updates rather than being averaged into one class gradient.

OLM-SGD is used as a full finetuning optimization technique to adapt the CLIP model. It updates the original parameters inside the selected Transformer blocks of both encoders. Parameters outside those blocks remain frozen. The prompt strings remain fixed, but their prototypes depend on the current text-encoder weights. This permits adaptation of both modalities without adding trainable modules. The adapted model is evaluated on disjoint query images from the same C classes. The query set is the test set of each dataset. These images do not contribute to parameter updates.

B EXTENDED RELATED WORK

This appendix expands Section 2 around the adaptation mechanisms evaluated in this work and the closest antecedents of our gradient-merging construction. The central distinction is between restricting how model parameters can change and structuring individual gradients before they are combined. Appendix D specifies the evaluated implementations and their training schedules.

B.1 Few-Shot Adaptation of Vision–Language Models

Prompt learning. CoOp (Zhou et al., 2022) replaces handcrafted context words with continuous trainable vectors. The pretrained CLIP weights remain frozen, but the learned prompts alter the text encoder's activations and resulting class representations. KgCoOp (Yao et al., 2023) constrains this adaptation by penalizing the discrepancy between learned and handcrafted text embeddings. ProGrad (Zhu et al., 2023) intervenes in optimization: a reference gradient derived from agreement with zero-shot predictions identifies conflicting components of the supervised prompt gradient. It projects out these components when the gradients conflict. Thus, knowledge preservation can operate through either an embedding penalty or a gradient constraint.

These methods establish that reliable adaptation requires more than fitting the support labels. Their intervention nevertheless remains in the prompt space. OLM-SGD addresses a different question: how to construct updates when the original residual-block parameters of both encoders are trainable. It combines gradients of class-distinct support examples rather than comparing a prompt gradient with a pretrained-reference gradient. The update therefore uses relationships among supervised examples within the current batch.

Adapters, caches, and projection tuning. CLIP-Adapter (Gao et al., 2024) adds a bottleneck transformation and mixes its output with pretrained features through a residual connection. Our Adapter baseline uses the image-feature configuration with frozen CLIP encoders. TIP-Adapter (Zhang et al., 2022) initializes a key–value cache from support-image features and labels. The evaluated TIP-Adapter-F variant subsequently learns the cache keys and combines cache predictions with zero-shot CLIP scores. It is therefore distinct from the training-free TIP-Adapter variant. Neither evaluated method updates the encoder residual blocks.

ProLIP (Fahes et al., 2026) fine-tunes the visual projection into CLIP's joint embedding space and penalizes its deviation from the pretrained projection. Its regularization explicitly addresses few-shot optimization reliability. OLM-SGD extends the scope of direct adaptation to the residual blocks of both encoders and controls the update through example-wise spectral processing. This moves the optimization intervention from a regularized output projection to gradient construction throughout the trainable residual blocks.

Low-rank parameterizations. LoRA (Hu et al., 2022) represents a weight correction as a product of two trainable low-rank factors while freezing the pretrained matrix. PiSSA (Meng et al., 2024) instead initializes trainable factors from principal singular components of the pretrained matrix and freezes the complementary residual. Its trainable component is low-rank, while its change from initialization is a difference of two low-rank matrices. DoRA (Liu et al., 2024) decomposes a weight matrix into magnitude and direction. It learns magnitudes and applies low-rank adaptation to the directional component. Magnitude adaptation gives DoRA additional freedom beyond a plain low-rank correction.

Our comparison uses the stated rank-two configurations in the attention projections of both encoders. OLM-SGD uses rank to approximate each current example gradient, with new singular factors computed at each iteration. The parameter tensors themselves are not represented by trainable low-rank factors. Successive updates can therefore modify different subspaces, and the cumulative parameter change need not satisfy a fixed low-rank constraint. The rank therefore constrains individual gradient approximations rather than the trainable parameterization. The accuracy comparison with LoRA and PiSSA in Table 1 supports retaining broader parameter flexibility under the reported protocol.

Sparse parameter and update selection. SAFT (Nguyen et al., 2025) uses gradient information to identify a small subset of pretrained parameters for adaptation. The selected support remains fixed. Our evaluated configuration selects the largest initial gradient magnitudes within each tensor and applies sparse Adam updates on that support. SO (Mrabah et al., 2025) instead allows the support to change. Its sparse Adam-style construction samples candidate gradient coordinates and retains large first-moment entries while maintaining sparse moment states. These are distinct mechanisms: SAFT restricts the accessible coordinates once, whereas SO repeatedly reallocates the update support.

Both methods directly adapt original parameters, but they control adaptation through coordinate selection. OLM-SGD preserves tensor structure in each individual gradient, truncates its singular components, and transforms the concatenated factors before reconstruction. It does not impose a sparse coordinate mask or use adaptive-moment histories. This gives a complementary way to control adaptation. On 1-shot ViT-B/16, OLM-SGD has higher epoch-300 dataset-average accuracy and smaller mean peak-to-final drops than LoRA, SAFT, and SO at each of the four tested learning rates (Appendix F.3). These observations support spectral gradient merging as an effective alternative under the shared long-horizon diagnostic.

Frozen-feature statistical and kernel classifiers. GDA (Wang et al., 2024) estimates class means and a shared covariance model from CLIP support features. A regularized precision estimate yields discriminant scores that are combined with zero-shot predictions. ProKeR (Bendou et al., 2025) derives a proximal kernel-ridge solution that balances support-label fitting with proximity to the initial predictor. Both construct task-specific classifiers without iterative encoder training. They provide classifier-based alternatives to trainable PEFT modules.

Their restriction is specific: supervision changes the decision rule over frozen features, not the feature extractor. OLM-SGD uses the same scarce labels to update the original encoder residual blocks. GDA is a strong competitor at higher shot counts, as Table 1 shows. The higher dataset averages of OLM-SGD therefore support direct representation adaptation against competitive classifier-only alternatives. They do not establish dominance on every dataset. Appendix F.1 provides the complete comparison with all 13 adaptation baselines.

B.2 Model Merging

Combining trained solutions. Model soups (Wortsman et al., 2022a) average fine-tuned checkpoints to improve a single resulting model. Task Arithmetic (Ilharco et al., 2023) represents adaptation by a task vector, defined as the parameter difference between a fine-tuned model and its shared initialization. Weighted combinations of these vectors combine learned capabilities. TIES-Merging (Yadav et al., 2023) trims small changes, elects a shared sign, and averages sign-compatible entries. These methods operate after the individual training trajectories have produced model parameters. Their objective is to combine trained solutions, whereas OLM-SGD determines the updates along a single training trajectory.

The direct TSV-M antecedent. TSV-M (Gargiulo et al., 2025) preserves the matrix structure of task vectors. Its construction first truncates the SVD of each task matrix. It concatenates the retained left factors and the retained right factors separately. An SVD-based transformation replaces each concatenated factor by the product of its left singular-vector matrix and the transpose of its right singular-vector matrix. Reconstruction then uses these transformed factors together with the retained task singular values. OLM-SGD inherits this two-sided singular-factor construction, including truncation, factor orthogonalization, and singular-value-weighted reconstruction. This shared construction provides the starting point for our online optimization formulation.

OLM-SGD changes the inputs and the role of that construction. Its inputs are individual-example gradients evaluated at the same current parameters, not differences between independently trained checkpoints and a pretrained model. Class-unique minibatching ensures distinct labels within a merge while retaining the full multiclass loss for every example. Different shots of the same class enter different updates. The retained singular values are divided by the realized batch size B_t, so the factors before the polar transformations reconstruct the mean of the truncated gradients. After applying the resulting update, all factors are recomputed at the new iterate. This defines a nonlinear optimizer rather than a one-time checkpoint combination.

The rank and shape conventions also serve the online formulation. OLM-SGD caps each gradient rank by its tensor dimensions, as specified by r_W in equation 4. The concatenated factors can be tall, wide, or rank-deficient as B_t and the tensor dimensions vary. Section 3.2 specifies their rectangular polar factors. Column orthogonality holds in the tall case; a wide factor instead has orthonormal rows. The construction therefore does not assert mutually orthogonal class contributions for every tensor. Section 4 and Appendix E analyze the resulting retained-gradient alignment, energy control, and optimization conditions. This sampling–update formulation and its few-shot analysis and evaluation are the contributions beyond the spectral antecedent.

B.3 Structured Optimization

Averaging and adaptive preconditioning. SGD uses the arithmetic mean of the minibatch gradients. Adam (Kingma & Ba, 2015) maintains exponential moving averages of gradients and their coordinate-wise squares to construct adaptive updates. Shampoo (Gupta et al., 2018) uses matrix-valued statistics along tensor dimensions to precondition gradients. These approaches process an already aggregated gradient and, for Adam and Shampoo, accumulate optimizer history. OLM-SGD keeps each example's tensor gradient available for a separate decomposition before any arithmetic aggregation. Its transient factors encode the current batch, and it maintains neither momentum nor second-moment buffers. Its purpose is to structure current class-distinct contributions rather than estimate a historical preconditioner.

Low-rank gradient and state compression. GaLore (Zhao et al., 2024) projects a layer's gradient into reduced coordinates, applies an optimizer in that space, and maps the update back to the original parameter dimensions. Periodically refreshed projection bases allow the optimized subspace to change. Thus, GaLore is not equivalent to a fixed low-rank cumulative parameter correction. The evaluated configuration uses Adam-style moments after gradient averaging (Appendix H.4). OLM-SGD decomposes the individual gradients first, then transforms both concatenated singular-factor matrices for the current batch. It neither compresses a persistent moment state nor reconstructs an update from one projected minibatch mean. The comparison isolates two uses of low-rank structure: reducing the coordinates of an aggregated optimizer update and preserving separate example contributions until the merge.

Polar matrix updates. Muon uses a matrix momentum update followed by an approximate polar transformation (Jordan, 2024). In its idealized SVD form, this replaces the momentum singular values by unit coefficients before the method's update scaling (Vasudeva et al., 2026). OLM-SGD applies polar transformations to different objects: the concatenated left and right singular factors of separately truncated example gradients. It retains their scaled singular values as reconstruction coefficients. These coefficients need not equal the final update's singular values when the factors are wide. Hence, polar transformations alone do not identify the two optimizers. Their inputs, aggregation order, spectral weighting, and use of history differ. Section 4 analyzes the geometry and optimization conditions of the resulting OLM-SGD update.

Multi-task gradient combination. PCGrad (Yu et al., 2020) projects away gradient components with conflicting pairwise inner products. CAGrad (Liu et al., 2021) balances progress on the average loss with worst-task local improvement. Nash-MTL (Navon et al., 2022) obtains combination weights through a bargaining formulation. These methods use task-level gradients to select a joint update. Their motivations explain why arithmetic averaging need not be the only combination rule, but they do not imply that every overlap among our class-distinct gradients is harmful.

Aligned-MTL (Senushkin et al., 2023) is a closer spectral comparison. It forms a matrix whose columns are vectorized task gradients and equalizes its nonzero singular values to the smallest positive singular value before applying task weights. OLM-SGD instead preserves the matrix structure of each parameter tensor within each example gradient. It truncates these matrices individually and transforms the concatenated left and right factors separately. The retained individual singular values remain reconstruction coefficients rather than being replaced by a common value. The supervised contributions are also different: OLM-SGD uses examples with distinct labels from one classification objective, not separate prediction tasks. No simultaneous-descent guarantee for all represented classes is required by our method description.

Author-designed controls. The controls in Appendix C.2 isolate decisions within the proposed optimizer. Class-unique SGD retains the sampler and uses arithmetic averaging. CTF-SGD uses randomized minibatches and an estimated cross-term-free energy scale to shrink the SGD update while preserving its direction. Removing polar reconstruction retains truncation but averages the truncated gradients. Class-mean merging averages shots within each class before the merge. These are explicitly defined experimental controls, not separate cited algorithms or additional proposed methods. Appendix H reports the evaluated optimizer and component comparisons.

B.4 Robust Fine-Tuning

WiSE-FT (Wortsman et al., 2022b), labeled WiSE in our tables, interpolates pretrained and fine-tuned parameters to balance adaptation with pretrained robustness. VRF (Zhu et al., 2024) instead combines zero-shot and fine-tuned predictions using an input-dependent weight informed by distances to examples on which fine-tuning repairs a zero-shot error. CLIPood (Shu et al., 2023) modifies the training objective through margin metric softmax and averages the parameter trajectory using a Beta moving average. Its original formulation tunes the image encoder with the text encoder frozen. Our evaluated implementation adapts the residual blocks of both encoders, as specified in Appendix D.

These methods preserve pretrained behavior through prediction combination, parameter interpolation, or an altered objective and temporal averaging. They do not determine how class-distinct instantaneous tensor gradients should be merged. OLM-SGD addresses that update construction directly, without requiring those additional mechanisms. Appendix G evaluates Adam and SGD variants of each robust baseline; these variants are experimental configurations of the same published method. The strong dependence on the base optimizer in those results motivates examining the update rule itself. Under the reported few-shot schedules, OLM-SGD has higher dataset-average accuracy in 14 of 15 settings. This supports improving the update rule alongside the prediction- and parameter-combination mechanisms studied in robust fine-tuning.

C Algorithms

This appendix specifies the optimizers used to examine the components of OLM-SGD. Each comparison addresses a different question. The SGD controls separate class-unique sampling from gradient merging. GaLoRE operates low-rank projections after gradient averaging, while OLM-SGD performs low-rank projections before class-specific gradient merging. CTF-SGD (Cross-Term-Free SGD) examines whether controlling the update magnitude can account for the behavior of OLM-SGD. The remaining variants either remove polar reconstruction or replace individual-example gradients using our class-unique batching algorithm with class-specific gradient means. OLM-SGD denotes the proposed method, which leverages the class-unique batching algorithm. The detailed name, OLM-SGD (Class-Unique Mini-Batching), denotes the same algorithm.

Randomized and class-unique mini-batching. A randomized epoch uniformly shuffles the N support indices and partitions them into consecutive minibatches of capacity B. It retains the final incomplete minibatch and imposes no label constraint. In contrast, Algorithm 1 shuffles each class's K examples and visits them in K rounds. Each round shuffles the classes and partitions them into groups of at most B. It selects one example from every represented class. Both schemes visit each support example once per epoch. Their update counts are $\lceil CK/B \rceil$ and $K\lceil C/B \rceil$, respectively. These counts can differ when B does not divide C. At $K = 1$, the two schemes have the same distribution because every support label is unique.

In every algorithm, all gradient contributions to an update are evaluated at the same Θ_t. A class-unique minibatch still uses the full C-way classification loss. Neither a single-example backward pass nor a microbatch split reduces the classifier to the labels present in that group. Learning rates are denoted by the η_t, and E is the number of complete support-set passes.

C.1 Complete OLM-SGD Procedure

Algorithms 1 and 2 specify the proposed method. The former produces one complete support-set pass. The latter evaluates individual gradients at unchanged parameters and performs one update after every merge. The rank and polar conventions are those of Section 3.

Indexed sampler definition. Let $\mathcal{I}_c$ contain the K support indices with label c. After independently shuffling these lists, let π_k be the class permutation for round k. For $j = 1, \ldots, \lceil C/B \rceil$, define

$$\begin{aligned} \mathcal{C}_{k,j} &= \{\pi_k[(j-1)B+1], \ldots, \pi_k[\min(jB, C)]\}, \\ \mathcal{B}_{k,j} &= \{(x_{\mathcal{I}_c[k]}, c) : c \in \mathcal{C}_{k,j}\}. \end{aligned} \tag{22}$$

Square brackets select entries of ordered lists. Each batch is processed in the order inherited from π_k. These batches become successive $\mathcal{B}_t$ and satisfy

$$y_{i_{t,b}} \neq y_{i_{t,b'}} \quad \text{for all distinct } b, b' \in \{1, \dots, B_t\}. \tag{23}$$

There are $K\lceil C/B \rceil$ batches per epoch, including each round's incomplete final batch. The implementation pops elements from shuffled class lists. Indexing a reversed shuffled list gives exactly the description in (22).

Algorithm 1 Class-unique epoch iterator for OLM-SGD

Require: Balanced support set $\mathcal{D}_{\text{sup}}$ with C classes and K examples per class; capacity $1 \leq B \leq C$
Ensure: Ordered batches satisfying (23), with every support index used once

```
for c = 1, ..., C do
    Form I_c ← [i : y_i = c]; verify |I_c| = K
    Replace I_c by an independent uniform permutation of its entries
end for
for k = 1, ..., K do
    Draw a fresh uniform permutation π_k of {1, ..., C}
    for j = 1, ..., ⌈C/B⌉ do
        Form C_{k,j} and B_{k,j} using (22)
        yield B_{k,j} in permutation order, retaining an incomplete final chunk
    end for
end for
```

Tensor identities and storage. For every $W \in \mathcal{W}$, the effective rank is initialized separately:

$$r_W = \min\{r, m_W, n_W\}. \tag{24}$$

The entry-preserving reshapes imply $d = \sum_{W \in \mathcal{W}} m_W n_W$. For a fixed tensor, write $u_{t,a}^{(b)} \in \mathbb{R}^{m_W}$ and $v_{t,a}^{(b)} \in \mathbb{R}^{n_W}$ for columns of its retained factors. Then

$$\Pi_{r_W}(G_t^{(b)}) = \sum_{a=1}^{r_W} \sigma_{t,a}^{(b)} u_{t,a}^{(b)} v_{t,a}^{(b)\top}. \tag{25}$$

The polar factors used in (10) are explicitly

$$U_{t,\text{pol}} = \text{polar}(U_t) \in \mathbb{R}^{m_W \times R_t}, \qquad V_{t,\text{pol}} = \text{polar}(V_t) \in \mathbb{R}^{n_W \times R_t}. \tag{26}$$

The diagonal construction $\text{diag}(s_t)$ has size $R_t \times R_t$. Restoring the tensor superscripts, the assembled update in (11) is

$$\widehat{g}_t = \text{concat}_{W \in \mathcal{W}} \text{vec}\Big(\text{mat}_W^{-1}(\widehat{G}_t^W)\Big) \in \mathbb{R}^d. \tag{27}$$

Here, vec lists entries in the fixed parameter order and concat joins the resulting vectors. Throughout, $\|\cdot\|_F$ is the Frobenius norm; $\|\cdot\|_2$ is Euclidean for vectors and spectral for matrices. Both polar factors have spectral norm one, so

$$\|\widehat{G}_t^W\|_F^2 \leq \|s_t^W\|_2^2 = \frac{1}{B_t^2} \sum_{b=1}^{B_t} \|\Pi_{r_W}(G_t^{(b),W})\|_F^2. \tag{28}$$

Equality holds if both concatenations are tall, including rank-deficient inputs under the completed reduced-SVD convention. For a wide factor, orthogonality concerns its rows and the expression remains an upper bound. Appendix E gives the complete geometry and its optimization consequences.

Numerical decomposition and computational cost. The implementation promotes gradients to FP32 for decomposition and reconstruction. In exact mode, it computes a reduced SVD and takes its leading r_W components. If the randomized mode is enabled, it uses `torch.svd_lowrank` with `q`=r_W and two subspace iterations when $r_W < \min(m_W, n_W)$ and $\max(m_W, n_W) > 32$. It uses exact reduced SVD for the remaining tensors. The best-rank statement in (5) concerns the exact mode. With numerical approximation, the returned triplet replaces that equation's exact factors,

Algorithm 2 OLM-SGD: complete class-unique training procedure

Require: Θ_0, balanced support $\mathcal{D}_{\mathrm{sup}}$, epochs $E \geq 1$, capacity $1 \leq B \leq C$, integer rank $r \geq 1$, learning rates $\{\eta_t\}_{t=0}^{EK\lceil C/B\rceil-1}$

Require: Ordered tensors $\mathcal{W}$ connected to each individual loss, reshape maps $\{\mathrm{mat}_W\}$, configured exact or randomized gradient-SVD mode

Ensure: Parameters after E complete passes, using $T = EK\lceil C/B\rceil$ updates

Initialize parameters to Θ_0 and set $t \leftarrow 0$
for each $W \in \mathcal{W}$ **do**
    Determine (m_W, n_W) from mat_W; set r_W using (24)
end for
for $e = 1, \ldots, E$ **do**
    Initialize a fresh epoch iterator by calling Algorithm 1 with $(\mathcal{D}_{\mathrm{sup}}, B)$
    for each ordered batch $\mathcal{B}_t$ yielded by that iterator **do**
        Set $B_t \leftarrow |\mathcal{B}_t|$ and record its support indices $i_{t,1}, \ldots, i_{t,B_t}$
        Evaluate full-C losses $\ell_{i_{t,b}}(\Theta_t)$ using (21); hold all parameters fixed until line 27
        for each $W \in \mathcal{W}$ **do**
            Initialize empty lists for its retained left factors, singular values, and right factors
        end for
        for $b = 1, \ldots, B_t$ **do**
            Clear previous autodiff gradients and compute $g_t^{(b)}$ using (2)
            for each $W \in \mathcal{W}$ **do**
                Form $G_t^{(b),W}$ using (4)
                Compute retained triplet $(U_t^{(b),W}, \Sigma_t^{(b),W}, V_t^{(b),W})$ using (5), or its configured numerical approximation in Appendix C.1
                Append the triplet to W's three factor lists, preserving batch order
            end for
            Release the dense gradient; retain the loss graph only if another backward pass needs it
        end for
        for each $W \in \mathcal{W}$ **do**
            Concatenate factors using (6); scale singular values by $1/B_t$ using (7)
            Compute both reduced-SVD polar factors using (9) and (26)
            Reconstruct $\widehat{G}_t^W$ using (10)
        end for
        Assemble $\widehat{g}_t$ using (27); apply one update using (11)
        Discard factor lists, clear gradients, and set $t \leftarrow t + 1$
    end for
end for
return Θ_T

and retained energies and residuals must be computed from the returned reconstruction. The polar operation always uses a reduced SVD and retains all of its columns, including zero-singular-value completions. No singular-value threshold or ridge term is applied.

The specified tensor collection is connected to every individual loss. In general-purpose use, the implementation skips an absent autodiff gradient (`None`) but retains an explicitly zero gradient. These cases are not interchangeable under completed polar reconstruction: an absent contribution supplies no columns, whereas a zero gradient supplies zero coefficients with SVD-selected factor columns. If only some losses supply a tensor gradient, its concatenation contains only those contributions and is still normalized by the full B_t. Such partial coverage falls outside the fixed-width formula $R_t = r_W B_t$ and must be recorded separately.

For each connected trainable tensor, an epoch requires CK individual-gradient decompositions and $2K\lceil C/B \rceil$ polar decompositions. Individual gradients can be compressed as they are computed. The full collection of dense example gradients need not coexist. The method stores retained factors for the current update but no optimizer moments. It introduces no inference-time modules.

C.2 Control Optimizers

Adam (Randomized Mini-Batching). Adam (Kingma & Ba, 2015) first averages the example gradients. It then maintains exponential moving averages of this mean gradient and its elementwise square. Let $\mathsf{Mom}^{(1)}_{t,\mathrm{Adam}}$ and $\mathsf{Mom}^{(2)}_{t,\mathrm{Adam}}$ denote these two buffers before update t. They are vectors in $\mathbb{R}^d$. The coefficients $\beta_1, \beta_2 \in [0, 1)$ control their decay, and $\varepsilon_{\mathrm{Adam}} > 0$ stabilize the denominator. We use the default PyTorch Adam values, namely $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon_{\mathrm{Adam}} = 10^{-8}$.

This baseline tests whether standard adaptive moment estimation is sufficient to provide robust few-shot fine-tuning without modifying the minibatch gradient geometry. Adam first forms the minibatch mean gradient and then updates its historical first- and second-moment estimates from this aggregated gradient. Consequently, the individual-example gradients are combined before adaptive scaling. In particular, the elementwise square of the minibatch mean differs from the mean of the elementwise squared individual gradients. Therefore, Adam does not preserve the individual gradient contributions that determine the OLM-SGD update. It also performs neither per-example low-rank approximation nor polar reconstruction. The comparison with Adam consequently evaluates whether conventional stateful adaptive preconditioning can reproduce the behavior of OLM-SGD, rather than isolating a single component of its gradient-merging procedure.

Algorithm 3 Adam (Randomized Mini-Batching)

Require: Θ_0, $\mathcal{D}_{\mathrm{sup}}$, epochs E, capacity B, $\{\eta_t\}$, $\beta_1, \beta_2, \varepsilon_{\mathrm{Adam}} > 0$

$t \leftarrow 0$; $\mathsf{Mom}^{(1)}_{0,\mathrm{Adam}} \leftarrow 0$; $\mathsf{Mom}^{(2)}_{0,\mathrm{Adam}} \leftarrow 0$
for $e = 1, \ldots, E$ **do**
  for each $\mathcal{B}_t$ in a freshly randomized epoch **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    $g_t \leftarrow \nabla_\Theta \big[B_t^{-1} \sum_{i \in \mathcal{B}_t} \ell_i(\Theta_t)\big]$
    $\mathsf{Mom}^{(1)}_{t+1,\mathrm{Adam}} \leftarrow \beta_1 \mathsf{Mom}^{(1)}_{t,\mathrm{Adam}} + (1 - \beta_1) g_t$
    $\mathsf{Mom}^{(2)}_{t+1,\mathrm{Adam}} \leftarrow \beta_2 \mathsf{Mom}^{(2)}_{t,\mathrm{Adam}} + (1 - \beta_2) g_t^2$
    $\widehat{\mathsf{Mom}}^{(1)}_{t+1,\mathrm{Adam}} \leftarrow \mathsf{Mom}^{(1)}_{t+1,\mathrm{Adam}} / (1 - \beta_1^{t+1})$
    $\widehat{\mathsf{Mom}}^{(2)}_{t+1,\mathrm{Adam}} \leftarrow \mathsf{Mom}^{(2)}_{t+1,\mathrm{Adam}} / (1 - \beta_2^{t+1})$
    $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \dfrac{\widehat{\mathsf{Mom}}^{(1)}_{t+1,\mathrm{Adam}}}{\sqrt{\widehat{\mathsf{Mom}}^{(2)}_{t+1,\mathrm{Adam}}} + \varepsilon_{\mathrm{Adam}}}$
    $t \leftarrow t + 1$
  end for
end for

SGD (Randomized Mini-Batching). This baseline applies the ordinary minibatch mean directly, without momentum or adaptive moments. Its update is

$$\Theta_{t+1} = \Theta_t - \eta_t g_t. \tag{29}$$

It provides the reference for asking whether OLM-SGD improves on direct full finetuning without changing the model architecture. A single backward pass through the mean loss computes g_t.

Two differences separate this baseline from OLM-SGD: the sampling rule and the merge. Randomized minibatches can contain repeated labels when $K > 1$. SGD does not process the individual-example gradients separately. It updates the parameters using only their arithmetic mean g_t. Thus, this comparison alone cannot attribute the improvement in test accuracy, the long-horizon stability, and the low-sensitivity to the learning rate specifically to class-unique sampling, truncation, or polar reconstruction. The next control holds the sampler fixed and removes the spectral merge (i.e., class-specific low-rank truncation and polar reconstruction).

Algorithm 4 SGD (Randomized Mini-Batching)

Require: Θ_0, $\mathcal{D}_{\mathrm{sup}}$, epochs E, capacity B, learning rates $\{\eta_t\}$
$t \leftarrow 0$
for $e = 1, \ldots, E$ **do**
  for each $\mathcal{B}_t$ in a freshly randomized epoch **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    $g_t \leftarrow \nabla_\Theta \big[B_t^{-1} \sum_{i \in \mathcal{B}_t} \ell_i(\Theta_t)\big]$
    $\Theta_{t+1} \leftarrow \Theta_t - \eta_t g_t$
    $t \leftarrow t + 1$
  end for
end for

SGD (Class-Unique Mini-Batching). This control uses the class-unique sampling rule of OLM-SGD but keeps the SGD update in (29). It matches the classes, examples, and actual minibatch sizes used by OLM-SGD when the sampler realizations are paired. It does not compute singular factors. This variant leverages a single backward pass for every minibatch.

The comparison with randomized SGD measures the effect of the batching rule. The comparison with OLM-SGD asks whether sampling alone explains its behavior. A remaining difference under matched batches must involve the spectral update. Class uniqueness does not imply an unconditional variance reduction relative to every without-replacement sampler. Appendix E.9 makes the sampling assumptions explicit. In particular, this control is *not* a K-shot class-mean optimizer. Different shots of one class still enter different updates.

Algorithm 5 SGD (Class-Unique Mini-Batching)

Require: Θ_0, $\mathcal{D}_{\mathrm{sup}}$, $\{\mathcal{I}_c\}_{c=1}^{C}$, epochs E, capacity $B \le C$, $\{\eta_t\}$
$t \leftarrow 0$
for $e = 1, \ldots, E$ **do**
  for each $\mathcal{B}_t$ yielded by Algorithm 1 **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    $g_t \leftarrow \nabla_\Theta \big[B_t^{-1} \sum_{i \in \mathcal{B}_t} \ell_i(\Theta_t)\big]$
    $\Theta_{t+1} \leftarrow \Theta_t - \eta_t g_t$
    $t \leftarrow t + 1$
  end for
end for

GaLoRE (Randomized Mini-Batching). GaLoRE (Zhao et al., 2024) first computes the minibatch mean gradient and then projects this aggregated gradient into a low-rank subspace. The evaluated implementation uses a one-sided low-rank projection and performs an Adam-style update in the projected coordinates. The resulting direction is projected back to the original parameter space before the model update. For a controlled comparison, GaLoRE adapts the same trainable parameter collection $\mathcal{W}$ as OLM-SGD. Thus, the two methods differ in how they process the gradients rather than in which model parameters they adapt. GaLoRE applies low-rank projection after gradient averaging, whereas OLM-SGD applies low-rank approximation to each individual-example gradient before merging them.

Let $r_{\mathrm{GaL}} \ge 1$ denote the GaLoRE projection-rank budget and define

$$r_W^{\mathrm{GaL}} = \min\{r_{\mathrm{GaL}}, m_W, n_W\}, \qquad W \in \mathcal{W}. \tag{30}$$

The fixed reshaping operator mat_W defined in Section 3.1 is used for every $W \in \mathcal{W}$, so the projection is defined over the complete shared trainable parameter set.

This rank constrains the projection of the already-averaged gradient. In contrast, the OLM-SGD rank r_W is applied separately to each individual-example gradient before merging. Hence, the two rank parameters act at different stages of the optimization procedure.

For every $W \in \mathcal{W}$, the projection basis $Q^W_{t,\mathrm{GaL}}$ is refreshed at $t = 0$ and every $T_{\mathrm{proj}} \geq 1$ updates. When $m_W \geq n_W$, the basis contains the leading r^{GaL}_W right singular vectors of $\mathrm{mat}_W(g^W_t)$. When $m_W < n_W$, it contains the corresponding leading left singular vectors. Between projection refreshes, the previous basis is reused. The projected gradient is

$$G^W_{t,\mathrm{proj}} = \begin{cases} \mathrm{mat}_W(g^W_t) Q^W_{t,\mathrm{GaL}}, & m_W \geq n_W, \\ (Q^W_{t,\mathrm{GaL}})^\top \mathrm{mat}_W(g^W_t), & m_W < n_W. \end{cases} \tag{31}$$

Accordingly, $G^W_{t,\mathrm{proj}}$ has shape $m_W \times r^{\mathrm{GaL}}_W$ when $m_W \geq n_W$, and $r^{\mathrm{GaL}}_W \times n_W$ otherwise.

GaLoRE maintains first- and second-moment buffers, $\mathsf{Mom}^{(1),W}_{t,\mathrm{GaL}}$ and $\mathsf{Mom}^{(2),W}_{t,\mathrm{GaL}}$, in these low-dimensional coordinates. These buffers are retained when the projection basis is refreshed. We use $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon_{\mathrm{GaL}} = 10^{-6}$, following the evaluated GaLoRE implementation. The Adam-preconditioned direction in the projected coordinates is

$$D^W_t = \frac{\sqrt{1-\beta_2^{t+1}}}{1-\beta_1^{t+1}} \frac{\mathsf{Mom}^{(1),W}_{t+1,\mathrm{GaL}}}{\sqrt{\mathsf{Mom}^{(2),W}_{t+1,\mathrm{GaL}}} + \varepsilon_{\mathrm{GaL}}}. \tag{32}$$

Before updating W, this low-dimensional direction is projected back to the original matrix space:

$$\Delta^W_t = \begin{cases} D^W_t (Q^W_{t,\mathrm{GaL}})^\top, & m_W \geq n_W, \\ Q^W_{t,\mathrm{GaL}} D^W_t, & m_W < n_W. \end{cases} \tag{33}$$

The parameter update is then

$$W_{t+1} = W_t - \eta_t \, \mathrm{mat}^{-1}_W(\Delta^W_t). \tag{34}$$

GaLoRE provides a direct comparison between low-rank processing after gradient averaging and the pre-merge low-rank processing used by OLM-SGD. For a data-dependent rank-r_W approximation, the two constructions are not identical in general:

$$\Pi_{r_W}\left(\frac{1}{B_t}\sum_{b=1}^{B_t} G^{(b),W}_t\right) \not\equiv \frac{1}{B_t}\sum_{b=1}^{B_t} \Pi_{r_W}\left(G^{(b),W}_t\right). \tag{35}$$

The distinction arises because the singular subspace is fitted to different gradients in the two cases. A fixed linear projector would instead commute with averaging. The right-hand side of Eq. (35) corresponds to averaging individually truncated gradients and therefore represents the pre-polar quantity $\widetilde{G}^W_t$, not the complete OLM-SGD update $\widehat{G}^W_t$. OLM-SGD subsequently orthogonalizes the retained factors before reconstruction. GaLoRE, by contrast, applies adaptive moment estimation in the projected coordinates and can reuse the projection basis across multiple updates.

Algorithm 6 GaLoRE (Randomized Mini-Batching): low-rank projection after gradient averaging

Require: $\Theta_0, \mathcal{D}_{\text{sup}}, E, B, \{\eta_t\}, \mathcal{W}, r_{\text{GaL}}, T_{\text{proj}}$
Require: $\beta_1, \beta_2 \in [0, 1), \varepsilon_{\text{GaL}} > 0$
$t \leftarrow 0$
$r_W^{\text{GaL}} \leftarrow \min\{r_{\text{GaL}}, m_W, n_W\}$ for every $W \in \mathcal{W}$
Initialize $\mathsf{Mom}_{0,\text{GaL}}^{(1),W}$ and $\mathsf{Mom}_{0,\text{GaL}}^{(2),W}$ to zero for every $W \in \mathcal{W}$
for $e = 1, \dots, E$ **do**
  for each $\mathcal{B}_t$ in a freshly randomized epoch **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    Compute the mean full-C-way loss over $\mathcal{B}_t$
    Compute its gradient g_t with respect to the tensors in $\mathcal{W}$
    for each $W \in \mathcal{W}$ **do**
      if $t \bmod T_{\text{proj}} = 0$ **then**
        Compute the reduced SVD of $\text{mat}_W(g_t^W)$
        Refresh $Q_{t,\text{GaL}}^W$ using the leading r_W^{GaL} singular vectors
      else
        $Q_{t,\text{GaL}}^W \leftarrow Q_{t-1,\text{GaL}}^W$
      end if
      Compute $G_{t,\text{proj}}^W$ using (31)
      $\mathsf{Mom}_{t+1,\text{GaL}}^{(1),W} \leftarrow \beta_1 \mathsf{Mom}_{t,\text{GaL}}^{(1),W} + (1-\beta_1) G_{t,\text{proj}}^W$
      $\mathsf{Mom}_{t+1,\text{GaL}}^{(2),W} \leftarrow \beta_2 \mathsf{Mom}_{t,\text{GaL}}^{(2),W} + (1-\beta_2)(G_{t,\text{proj}}^W)^2$
      Compute D_t^W using (32)
      Compute Δ_t^W using (33)
      $W_{t+1} \leftarrow W_t - \eta_t \, \text{mat}_W^{-1}(\Delta_t^W)$
    end for
    $t \leftarrow t + 1$
  end for
end for

CTF-SGD (Randomized Mini-Batching). Cross-Term-Free SGD (CTF-SGD) examines whether controlling the magnitude of the SGD update can recover part of the robustness of OLM-SGD without changing the SGD direction. The motivation follows from the different ways in which SGD and OLM-SGD combine gradient energy. For the ordinary minibatch mean

$$g_t = \frac{1}{B_t} \sum_{b=1}^{B_t} g_t^{(b)}, \tag{36}$$

its squared norm expands as

$$\|g_t\|_2^2 = \frac{1}{B_t^2} \sum_{b=1}^{B_t} \|g_t^{(b)}\|_2^2 + \frac{1}{B_t^2} \sum_{b \neq b'} \left\langle g_t^{(b)}, g_t^{(b')} \right\rangle. \tag{37}$$

The second term contains the pairwise interactions between individual gradients. When these gradients are positively aligned, the cross terms increase the norm of the minibatch mean. By contrast, OLM-SGD bounds update energy by retained self-energies. This bound is an equality when both concatenated polar factors have orthonormal columns. This observation motivates testing whether suppressing the corresponding excess magnitude is sufficient to explain part of OLM-SGD's stability.

CTF-SGD preserves the ordinary SGD direction and targets only its magnitude. Define the full-gradient self-energy

$$\mathcal{E}_t^{\text{CTF}} = \frac{1}{B_t^2} \sum_{b=1}^{B_t} \|g_t^{(b)}\|_2^2. \tag{38}$$

If the pairwise cross terms in (37) are positive, then $\|g_t\|_2^2$ can exceed $\mathcal{E}_t^{\text{CTF}}$. CTF-SGD therefore rescales g_t toward this cross-term-free energy while keeping its direction unchanged. It performs no low-rank truncation and no polar reconstruction. Hence, CTF-SGD isolates a norm-control mechanism rather than approximating the complete OLM-SGD direction.

The evaluated CTF-SGD baseline uses randomized mini-batching. At the beginning of each epoch, the support examples are randomly shuffled and divided into minibatches of capacity B. No class-unique constraint is imposed. Consequently, when $K > 1$, a minibatch can contain several examples from the same class. CTF-SGD then randomly partitions each realized minibatch into two nonempty microbatches. This second randomization is used only to estimate the cross-term-free energy and does not alter the underlying randomized minibatch sampler.

Computing (38) exactly requires access to all individual-example gradients. CTF-SGD avoids this cost by estimating the same quantity from two mean-loss gradients. For $B_t \geq 2$, uniformly permute the examples in the realized minibatch $\mathcal{B}_t$ and split them into $\mathcal{B}_{t,\mathrm{L}}$ and $\mathcal{B}_{t,\mathrm{R}}$, with

$$B_{t,\mathrm{L}} = \left\lfloor \frac{B_t}{2} \right\rfloor, \qquad B_{t,\mathrm{R}} = B_t - B_{t,\mathrm{L}}. \tag{39}$$

Both microbatch gradients are evaluated at the same parameter point Θ_t:

$$\begin{aligned} g_{t,\mathrm{L}} &= \nabla_\Theta \left[\frac{1}{B_{t,\mathrm{L}}} \sum_{i \in \mathcal{B}_{t,\mathrm{L}}} \ell_i(\Theta_t) \right], \\ g_{t,\mathrm{R}} &= \nabla_\Theta \left[\frac{1}{B_{t,\mathrm{R}}} \sum_{i \in \mathcal{B}_{t,\mathrm{R}}} \ell_i(\Theta_t) \right], \\ g_t &= \frac{B_{t,\mathrm{L}} g_{t,\mathrm{L}} + B_{t,\mathrm{R}} g_{t,\mathrm{R}}}{B_t}. \end{aligned} \tag{40}$$

The size-weighted reconstruction of g_t is necessary when B_t is odd. No parameter update is performed between the two backward passes.

The resulting estimator of the cross-term-free energy is

$$\widehat{\mathcal{E}}_t^{\mathrm{CTF}} = \frac{1}{B_t} \left[\|g_t\|_2^2 + \frac{(B_t - 1) B_{t,\mathrm{L}} B_{t,\mathrm{R}}}{B_t^2} \|g_{t,\mathrm{L}} - g_{t,\mathrm{R}}\|_2^2 \right]. \tag{41}$$

For fixed individual gradients and a uniform random split, the expectation of $\widehat{\mathcal{E}}_t^{\mathrm{CTF}}$ over the split equals $\mathcal{E}_t^{\mathrm{CTF}}$. In particular,

$$\mathbb{E}\left[\|g_{t,\mathrm{L}} - g_{t,\mathrm{R}}\|_2^2 \right] = \frac{B_t}{B_{t,\mathrm{L}} B_{t,\mathrm{R}} (B_t - 1)} \sum_{b=1}^{B_t} \|g_t^{(b)} - g_t\|_2^2, \tag{42}$$

which yields (38) after substitution into (41). The unbiasedness concerns the energy estimator itself. It does not imply that the square root or the subsequent clipping operation is unbiased.

Using numerical tolerance $\varepsilon_{\mathrm{CTF}} > 0$, CTF-SGD defines

$$\alpha_t^{\mathrm{CTF}} = \begin{cases} 1, & B_t = 1 \text{ or } \|g_t\|_2 \leq \varepsilon_{\mathrm{CTF}}, \\ \min\left\{ 1, \dfrac{\sqrt{\max\{\widehat{\mathcal{E}}_t^{\mathrm{CTF}}, 0\}}}{\|g_t\|_2 + \varepsilon_{\mathrm{CTF}}} \right\}, & \text{otherwise.} \end{cases} \tag{43}$$

The parameter update is

$$\Theta_{t+1} = \Theta_t - \eta_t \alpha_t^{\mathrm{CTF}} g_t. \tag{44}$$

For the evaluated baseline, the norms are computed globally over the complete trainable parameter collection $\mathcal{W}$. Thus, a single scalar α_t^{CTF} rescales the entire minibatch gradient. Since $0 \leq \alpha_t^{\mathrm{CTF}} \leq 1$, the update remains collinear with the ordinary SGD gradient and cannot have a larger norm. A singleton final minibatch uses an ordinary SGD update because it cannot be divided into two nonempty microbatches. When the minibatch mean is zero, the resulting parameter update also remains zero. We use $\varepsilon_{\mathrm{CTF}} = 10^{-12}$.

Accordingly, CTF-SGD is a scale-control baseline rather than an approximation to the complete OLM-SGD update. Its target uses the full individual-gradient energy, whereas OLM-SGD first retains only the leading singular components. The corresponding quantities satisfy

$$\|\widehat{g}_t\|_2^2 \leq \mathcal{Q}_t \leq \mathcal{E}_t^{\mathrm{CTF}}, \tag{45}$$

where $\mathcal{Q}_t$ is the retained-gradient energy budget defined in (49). The first inequality follows from the polar reconstruction, while the second follows from discarding gradient energy through low-rank truncation. Consequently, CTF-SGD does not generally reproduce either the norm or the direction of the OLM-SGD update. It instead tests whether suppressing constructive cross-term accumulation while preserving the SGD direction is sufficient to account for OLM-SGD's behavior.

Algorithm 7 CTF-SGD (Randomized Mini-Batching): global two-microbatch estimate

Require: $\Theta_0, \mathcal{D}_{\mathrm{sup}}, E, B, \{\eta_t\}, \varepsilon_{\mathrm{CTF}} > 0$
$t \leftarrow 0$
for $e = 1, \ldots, E$ **do**
  Randomly shuffle the support examples and partition them into minibatches of capacity B
  for each resulting minibatch $\mathcal{B}_t$ **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    if $B_t = 1$ **then**
      Compute g_t from the single loss
      $\alpha_t^{\mathrm{CTF}} \leftarrow 1$
    else
      Uniformly permute $\mathcal{B}_t$ and split it into $\mathcal{B}_{t,\mathrm{L}}$ and $\mathcal{B}_{t,\mathrm{R}}$
      $B_{t,\mathrm{L}} \leftarrow \lfloor B_t/2 \rfloor$; $B_{t,\mathrm{R}} \leftarrow B_t - B_{t,\mathrm{L}}$
      Compute $g_{t,\mathrm{L}}$ at Θ_t from the mean loss over $\mathcal{B}_{t,\mathrm{L}}$
      Compute $g_{t,\mathrm{R}}$ at the unchanged Θ_t from the mean loss over $\mathcal{B}_{t,\mathrm{R}}$
      Form the size-weighted mean gradient g_t using (40)
      Compute $\widehat{\mathcal{E}}_t^{\mathrm{CTF}}$ using (41)
      Compute α_t^{CTF} using (43)
    end if
    $\Theta_{t+1} \leftarrow \Theta_t - \eta_t \alpha_t^{\mathrm{CTF}} g_t$ ▷ one update after both microbatch gradients
    $t \leftarrow t + 1$
  end for
end for

OLM-SGD (Without Polar Reconstruction). This ablation retains class-unique mini-batching and applies the same rank-r_W truncation to every individual gradient. It then averages the truncated reconstructions directly. The update uses the already-defined prepolar vector:

$$\widetilde{G}_t^W = \frac{1}{B_t} \sum_{b=1}^{B_t} \Pi_{r_W}\left(G_t^{(b),W}\right), \qquad \Theta_{t+1} = \Theta_t - \eta_t \widetilde{g}_t. \tag{46}$$

It reconstructs $U_t^W \operatorname{diag}(s_t^W)(V_t^W)^\top$ without replacing either factor with its polar factor. It preserves the singular values, tensor reshaping, rank selection, and single division by B_t as OLM-SGD.

This control isolates polar reconstruction after fixing the sampler and retained components. It asks whether low-rank filtering alone explains the observed behavior. Direct averaging retains cross terms between truncated contributions, so $\|\widetilde{g}_t\|_2^2 \le \mathcal{Q}_t$ is not guaranteed. Positive alignment can enlarge its norm, whereas cancelation can reduce it. Removing polar reconstruction can therefore change both direction and magnitude. If every tensor retains its full matrix rank and the decompositions are exact, this ablation reduces to SGD with class-unique batches. Moreover, removing polar reconstruction is not equivalent to projecting the averaged gradient as in GaLoRE.

Algorithm 8 OLM-SGD (Without Polar Reconstruction)

Require: $\Theta_0, \mathcal{D}_{\text{sup}}, \{\mathcal{I}_c\}_{c=1}^{C}, E, B \le C, r, \{\eta_t\}, \mathcal{W}, \{\text{mat}_W\}$
$t \leftarrow 0$
$r_W \leftarrow \min\{r, m_W, n_W\}$ for every $W \in \mathcal{W}$
for $e = 1, \dots, E$ **do**
  for each $\mathcal{B}_t$ yielded by Algorithm 1 **do**
    $B_t \leftarrow |\mathcal{B}_t|$
    $\widetilde{G}_t^W \leftarrow 0$ for every $W \in \mathcal{W}$
    for $b = 1, \dots, B_t$ **do**
      Compute $g_t^{(b)}$ at the unchanged Θ_t
      for each $W \in \mathcal{W}$ **do**
        Form $G_t^{(b),W}$ and its rank-r_W factors using (5)
        $\widetilde{G}_t^W \leftarrow \widetilde{G}_t^W + \frac{1}{B_t} U_t^{(b),W} \Sigma_t^{(b),W} (V_t^{(b),W})^\top$
      end for
      Release the dense example gradient
    end for
    for each $W \in \mathcal{W}$ **do**
      $W_{t+1} \leftarrow W_t - \eta_t \, \text{mat}_W^{-1}(\widetilde{G}_t^W)$
    end for
    $t \leftarrow t + 1$
  end for
end for

OLM-SGD (Class-Mean Gradients). This variant changes the gradient contribution supplied to the spectral merge. Instead of using one current example from each selected class, it uses the mean gradient of *all K support examples* from that class. The class mean is $\mu_c(\Theta_t)$ from (83). One backward pass through the mean of the K losses computes this vector exactly.

Let $B_{\text{cls}} \le C$ be the capacity in complete classes. We use $B_{\text{cls}} = B$ when matching the number of contributions per merge. Let $\mathcal{C}_t^{\text{cm}}$ be the ordered group of selected classes and let $M_t^{\text{cm}} = |\mathcal{C}_t^{\text{cm}}| \le B_{\text{cls}}$. The capacity B_{cls} caps *classes*, not samples. The logical example group is $\mathcal{B}_t^{\text{cm}} = \bigcup_{c \in \mathcal{C}_t^{\text{cm}}} \mathcal{I}_c$, of size KM_t^{cm}. We do not denote its class count by B_t, which remains the number of individual examples in an ordinary or class-unique minibatch. For each selected class and tensor, compute

$$\Pi_{r_W}\left(\text{mat}_W(\mu_c^W(\Theta_t))\right) = U_{t,\text{cm}}^{(c),W} \Sigma_{t,\text{cm}}^{(c),W} (V_{t,\text{cm}}^{(c),W})^\top. \tag{47}$$

The superscript W on μ_c^W denotes its original-shaped tensor slice. The factors have the same shapes as in (5), with a complete class mean replacing an individual gradient.

Concatenate the left and right factors in the order of $\mathcal{C}_t^{\text{cm}}$ to obtain $U_{t,\text{cm}}^W$ and $V_{t,\text{cm}}^W$. Their width is $r_W M_t^{\text{cm}}$. Concatenate the diagonal entries of the $\Sigma_{t,\text{cm}}^{(c),W}$ in the same order and divide them by M_t^{cm} to obtain $s_{t,\text{cm}}^W$. The merged matrix is

$$\widehat{G}_{t,\text{cm}}^W = \text{polar}(U_{t,\text{cm}}^W) \, \text{diag}(s_{t,\text{cm}}^W) \, \text{polar}(V_{t,\text{cm}}^W)^\top. \tag{48}$$

The class loss already contributes the factor $1/K$. Reconstruction contributes $1/M_t^{\text{cm}}$, not an additional $1/K$ or $1/(KM_t^{\text{cm}})$.

This ablation asks whether averaging within a class before truncation gives a better contribution than preserving the shots separately across updates. Such averaging changes the matrix supplied to the SVD and can remove shot-specific components before the merge. The algorithm retains OLM's polar reconstruction and introduces no adaptive state. At $K = 1$, it recovers OLM-SGD under matched class groups and decomposition choices.

The update budget is also different when $K > 1$. Each epoch shuffles the classes once, partitions them into complete-class groups, and visits all CK examples. It makes $\lceil C/B_{\text{cls}} \rceil$ updates and C complete-class backward passes. OLM-SGD instead makes $K\lceil C/B \rceil$ updates and CK individual-example backward passes. With $B_{\text{cls}} = B$, equal epochs match example exposure but differ by a factor of K in optimization steps. We do not multiply the learning rate by K or repeat class groups to compensate. This comparison tests complete class-mean merging under its stated schedule, rather than an isolated variance change at a fixed update count.

Algorithm 9 OLM-SGD (Class-Mean Gradients)

Require: Θ_0, $\mathcal{D}_{\mathrm{sup}}$, $\{\mathcal{I}_c\}_{c=1}^{C}$, K, E, class capacity $B_{\mathrm{cls}} \le C$, r, $\{\eta_t\}$
Require: $\mathcal{W}$ and $\{\mathrm{mat}_W\}$
$t \leftarrow 0$
$r_W \leftarrow \min\{r, m_W, n_W\}$ for every $W \in \mathcal{W}$
for $e = 1, \dots, E$ **do**
  Shuffle the C classes and partition them into consecutive groups of at most B_{cls}
  for each ordered class group $\mathcal{C}_t^{\mathrm{cm}}$ **do**
    $M_t^{\mathrm{cm}} \leftarrow |\mathcal{C}_t^{\mathrm{cm}}|$
    for each $c \in \mathcal{C}_t^{\mathrm{cm}}$ **do**
      Compute $\mu_c(\Theta_t)$ by one backward pass through $K^{-1}\sum_{i\in\mathcal{I}_c}\ell_i(\Theta_t)$
      for each $W \in \mathcal{W}$ **do**
        Compute and retain $U_{t,\mathrm{cm}}^{(c),W}$, $\Sigma_{t,\mathrm{cm}}^{(c),W}$, $V_{t,\mathrm{cm}}^{(c),W}$ using (47)
      end for
      Release the dense class gradient
    end for
    for each $W \in \mathcal{W}$ **do**
      Concatenate the retained factors to form $U_{t,\mathrm{cm}}^{W}$ and $V_{t,\mathrm{cm}}^{W}$
      Form $s_{t,\mathrm{cm}}^{W}$ by concatenating the retained singular values and dividing by M_t^{cm}
      Compute $\widehat{G}_{t,\mathrm{cm}}^{W}$ using (48)
      $W_{t+1} \leftarrow W_t - \eta_t\,\mathrm{mat}_W^{-1}(\widehat{G}_{t,\mathrm{cm}}^{W})$
    end for
    $t \leftarrow t+1$
  end for
end for

D HYPERPARAMETERS

Shared image processing pipeline. For the standard few-shot comparisons, we use the same deterministic image preprocessing for all methods and datasets. Images are resized with bicubic interpolation and center-cropped to 224×224 pixels. They are converted to RGB tensors and normalized with the CLIP channel statistics. Data augmentation is not used in these standard comparisons. The extended ImageNet training trajectories are an exception: they retain random resized crops and horizontal flips. Evaluation remains deterministic. Appendix H states this exception explicitly for the trajectory analyses.

Table 2: Deterministic image preprocessing for the standard comparisons and evaluation. Extended ImageNet training trajectories retain random crops and horizontal flips, as stated in the text.

Operation	Setting
Resize	Shorter side 224, bicubic interpolation
Center crop	224×224
Color space	RGB
Tensor conversion	Pixel values scaled to $[0, 1]$
Data augmentation	None

We retain the notation of Section 3 and Appendix C. The learning rate is η_t, and E counts complete support-set epochs. The batch capacity is B, whereas B_t is the actual number of examples in an update. Randomized minibatching uses a capacity of 64 images. Class-unique minibatching uses $B = \min\{64, C\}$ and retains incomplete batches. Unless stated otherwise, the settings below refer to the standard few-shot comparisons with $K \in \{1, 2, 4, 8, 16\}$ and seeds 1, 2, and 3. The extended trajectory schedules used for stability and causal analyses are specified at the end of this appendix.

Optimization-based methods operate on the Transformer residual blocks of the indicated encoders. Parameters outside these blocks remain frozen unless the method explicitly adapts prompts, an adapter, or the final visual projection. GDA and ProKeR construct classifiers from frozen support features. They therefore use no optimizer, learning rate, or training epochs.

ProLIP. ProLIP (Fahes et al., 2026) updates only the final visual projection and freezes the remaining CLIP parameters. Its regularizer penalizes the squared Frobenius distance from the pretrained projection. We set its weight to $1/K$. This is a penalty relative to the pretrained projection.

Table 3: Hyperparameters for ProLIP.

Setting	Value
Trainable parameters	Final visual projection matrix
Optimizer	Adam
Learning rate	10^{-4}
Epochs (E)	300
Batch size	64
Projection-regularization weight	$1/K$
Adam moment coefficients	$(0.9,\ 0.999)$
Adam numerical epsilon	10^{-4}

Adapter. Adapter (Gao et al., 2024) learns a bottleneck transformation of the image features while keeping CLIP frozen. The bottleneck reduces the feature dimension by a factor of four. The adapted feature receives a fixed residual mixing weight of 0.2. This weight is not learned.

Table 4: Hyperparameters for Adapter.

Setting	Value
Optimizer	Adam
Learning rate	10^{-3}
Epochs (E)	30
Batch size	64
Bottleneck reduction factor	4
Adapted-feature mixing weight	0.2

SAFT. SAFT (Nguyen et al., 2025) configuration uses sparse Adam updates. For each trainable tensor, it selects a static support from the largest initial gradient magnitudes. The selected fraction is 0.001, or 0.1% of the coordinates. Only this support is updated during adaptation. The support is selected separately within each tensor rather than through a single model-wide ranking.

Table 5: Hyperparameters for SAFT.

Setting	Value
Optimizer	Sparse Adam
Learning rate	10^{-4}
Epochs (E)	30
Batch size	64
Density setting	0.001 (0.1%)
Support selection	Largest initial gradient magnitudes per tensor
Support refresh	None (static support)
Adam moment coefficients	$(0.9,\ 0.999)$
Adam numerical epsilon	10^{-8}
Adapted encoders	Vision and text

CoOp. CoOp (Zhou et al., 2022) learns 16 continuous context tokens while keeping the pretrained CLIP weights frozen. We optimize the context tokens with Adam for 200 epochs.

KgCoOp. KgCoOp (Yao et al., 2023) adds a knowledge-guided regularizer to prompt learning. We retain 16 context tokens and set the knowledge-guidance weight to 8.0. The optimizer is SGD.

ProGrad. ProGrad (Zhu et al., 2023) learns continuous context tokens while preserving the knowledge of zero-shot CLIP through gradient projection. When the cross-entropy and knowledge-

guidance gradients conflict, ProGrad removes the conflicting component. We use 16 learnable context tokens and set the gradient-projection coefficient to 1.0.

Table 6: Hyperparameters for the prompt-learning baselines. A dash denotes a setting that does not apply to the method.

Setting	CoOp	KgCoOp	ProGrad
Optimizer	Adam	SGD	Adam
Learning rate	10^{-4}	10^{-4}	10^{-4}
Epochs (E)	200	200	200
Batch size	64	64	64
Context tokens	16	16	16
Knowledge-guidance weight	–	8.0	–
Projection weight	–	–	1.0
Distillation temperature	–	–	1.0

TIP-Adapter. The method reported as TIP-Adapter (Zhang et al., 2022) uses the cache-fine-tuning configuration, also known as TIP-Adapter-F. The cache is initialized from normalized support features. Its keys are then optimized while the CLIP encoders remain frozen. We fix the cache-mixing weight to 0.27 and the cache-scaling weight to 20.

Table 7: Hyperparameters for the cache-fine-tuning configuration of TIP-Adapter.

Setting	Value
Optimizer	Adam
Learning rate	5×10^{-4}
Epochs (E)	10
Batch capacity	64
Cache-mixing weight	0.27
Cache-scaling weight	20.0

GDA. GDA (Wang et al., 2024) estimates class means and a shared precision matrix from normalized, frozen CLIP features. The precision matrix uses the Kubokawa–Srivastava empirical-Bayes ridge estimator. The resulting discriminant scores are combined with the zero-shot logits. The ensemble coefficient is set to 0.1. Support-feature extraction follows the deterministic pipeline above.

Table 8: Implementation settings for GDA. The coefficient-selection rule is shared across datasets and shot counts.

Setting	Value
Support-feature extraction passes	10
Precision estimator	Kubokawa–Srivastava empirical-Bayes ridge
Ensemble coefficient	0.1
Support-feature extraction batch size	64
Evaluation batch size	256
Numerical ridge	10^{-6}, only for a degenerate precision system
Optimization	Closed form

ProKeR. ProKeR (Bendou et al., 2025) solves proximal kernel ridge regression in closed form using normalized, frozen CLIP support features. We retain the ImageNet-transferred settings used in the official implementation. These settings depend only on K and remain unchanged across datasets and backbones. The support-feature and evaluation batch sizes are 64 and 256, respectively.

LoRA, DoRA, and PiSSA. LoRA (Hu et al., 2022) learns low-rank weight corrections. DoRA (Liu et al., 2024) additionally separates weight magnitude from direction. PiSSA (Meng et al., 2024) initializes its trainable factors from principal singular components of the pretrained weights. We use

Table 9: Shot-dependent ProKeR hyperparameters.

Shots (K)	Kernel setting	Proximal regularization
1	3.00	0.05
2	2.60	0.05
4	1.50	0.05
8	1.66	0.07
16	1.70	0.10

rank 2 and set the LoRA scaling hyperparameter $\alpha_{\mathrm{LoRA}} = 1$ for all three methods. The low-rank dimension is distinct from OLM-SGD's gradient rank r. Adaptation covers the attention projections in all Transformer blocks of both encoders.

Table 10: Shared hyperparameters for LoRA, DoRA, and PiSSA.

Setting	Value
Optimizer	Adam
Learning rate	10^{-4}
Epochs (E)	30
Batch capacity	64
Low-rank parameterization rank	2
LoRA Scaling coefficient (α_{LoRA})	1
Adapted encoders	Vision and text

SO. SO (Mrabah et al., 2025) sparsifies the gradient and Adam-style moment states on a periodically refreshed support. Its support update combines uniform random sampling of gradient coordinates with retention of large first-moment entries. We use an update density of 0.001 and refresh the support every 10 optimizer steps. Unlike SAFT, SO changes the selected coordinates during training.

Table 11: Hyperparameters for SO.

Setting	Value
Learning rate	10^{-4}
Epochs (E)	30
Batch capacity	64
Density setting	0.001 (0.1%)
Support update	Uniform random sampling and first-moment retention
Support-refresh period	10 optimizer steps
Moment coefficients	$(0.9, 0.999)$
Numerical epsilon	10^{-8}
Adapted encoders	Vision and text

Robust fine-tuning methods. We evaluate VRF (Zhu et al., 2024), WiSE (Wortsman et al., 2022b), and CLIPood (Shu et al., 2023) with both Adam and SGD. For each method, the method-specific hyperparameters are kept fixed across the two optimizers Adam and SGD. Only the base optimizer is changed. The shared training hyperparameters are listed in Table 12.

VRF. VRF (Zhu et al., 2024) combines the fine-tuned and zero-shot predictions using an input-dependent weight. We use sigmoid weighting with center 1.5 and scale 0.5. The number of nearest neighbors is set to 0.1% of the reference feature bank. Both the vision and text Transformer blocks are fine-tuned.

WiSE. WiSE (Wortsman et al., 2022b) interpolates pretrained and fine-tuned weights. The interpolation coefficient is fixed to 0.5, giving equal weight to the two models. Both the vision and text Transformer blocks are fine-tuned before interpolation.

Table 12: Shared training hyperparameters for VRF, WiSE, and CLIPood.

Setting	Value
Optimizer variants	Adam and SGD
Learning rate	10^{-4}
Epochs (E)	30
Batch size	64
Adam moment coefficients	$(0.9,\ 0.999)$
Adam numerical epsilon	10^{-8}
Adapted encoders	Vision and text

CLIPood. CLIPood (Shu et al., 2023) uses margin metric softmax and Beta moving average. We set the margin strength to 0.3 and the moving-average shape setting to 0.5. Both the vision and text Transformer blocks are fine-tuned.

Table 13: Method-specific hyperparameters for robust fine-tuning.

Method	Setting	Value
VRF	Sigmoid center	1.5
VRF	Sigmoid scale	0.5
VRF	Neighbor percentage	0.1%
WiSE	Weight-interpolation coefficient	0.5
CLIPood	Margin strength	0.3
CLIPood	Moving-average shape	0.5

Adam (Randomized Mini-Batching). Adam uses shuffled sample minibatches and the moment coefficients defined in Appendix C. We set $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon_{\mathrm{Adam}} = 10^{-8}$. All parameters inside the vision and text Transformer blocks are trainable.

SGD (Randomized Mini-Batching). SGD uses shuffled minibatches sampled from the support set. A minibatch may contain multiple samples from the same class. All parameters within the vision and text Transformer blocks are trainable. Momentum, Nesterov acceleration, and weight decay are disabled. The update follows the mean-gradient rule in (29).

SGD (Class-Unique Mini-Batching). This control retains the SGD update and replaces randomized minibatching with the class-unique iterator. We use $B = \min\{64, C\}$. The fixed-schedule results in Appendix H.2 use 30 epochs at 10^{-4}, following the evaluated class-unique SGD campaign. Randomized SGD also uses 30 epochs at 10^{-4}. The common-horizon comparisons match the learning rate and epoch count across both SGD controls and OLM-SGD. The class-unique SGD update introduces neither a spectral transformation nor adaptive optimizer state.

Table 14: Standard few-shot settings for the Adam and SGD controls.

Setting	Value
Learning rate (Adam and randomized SGD)	10^{-4}
Learning rate (class-unique SGD)	10^{-4}
Epochs (E)	30
Randomized batch size	64
Class-unique batch size (B)	$\min\{64, C\}$
Adam coefficients (β_1, β_2)	$(0.9,\ 0.999)$
Adam epsilon ($\varepsilon_{\mathrm{Adam}}$)	10^{-8}
Adapted encoders	Vision and text

GaLoRE (Randomized Mini-Batching). GaLoRE (Zhao et al., 2024) projects the full minibatch gradient into a low-rank subspace and maintains Adam-style moments in that subspace. We use $r_{\mathrm{GaL}} = 2$ and refresh the projection every $T_{\mathrm{proj}} = 10$ optimizer steps. GaLoRE adapts the same

parameter set as OLM-SGD, comprising all parameters within the Transformer residual blocks of both encoders.

Table 15: Hyperparameters for GaLoRE.

Setting	Value
Learning rate	10^{-4}
Epochs (E)	30
Randomized batch size	64
Projection rank (r_{GaL})	2
Projection-refresh period (T_{proj})	10 optimizer steps
Moment coefficients (β_1, β_2)	(0.9, 0.999)
Numerical epsilon ($\varepsilon_{\mathrm{GaL}}$)	10^{-6}
Adapted encoders	Vision and text

CTF-SGD (Randomized Mini-Batching). CTF-SGD uses randomized minibatches of capacity 64. Its fixed-schedule results use 30 epochs at 10^{-3}. The common-horizon experiments match the learning rate and epoch count across methods. CTF-SGD uses two microbatches to estimate the cross-term-free target magnitude. It applies one global shrink coefficient across all trainable tensors. The numerical tolerance is $\varepsilon_{\mathrm{CTF}} = 10^{-12}$. The shrink coefficient is computed from the two microbatch gradients at each optimizer step and is capped at one. Therefore, CTF-SGD cannot amplify the ordinary SGD update.

Table 16: Hyperparameters for CTF-SGD.

Setting	Value
Learning rate	10^{-3}
Epochs (E)	30
Randomized batch size	64
Microbatches per estimate	2 for $B_t \geq 2$
Numerical tolerance ($\varepsilon_{\mathrm{CTF}}$)	10^{-12}
Adapted encoders	Vision and text

OLM-SGD (Without Polar Reconstruction). This variant uses the same class-unique minibatching and rank $r = 3$ as OLM-SGD. For each individual-example gradient, it reconstructs the rank-r truncated approximation and directly averages these truncated gradients without applying the polar orthogonalization. Its learning rate, epoch budget, and decomposition settings match Table 17. Removing the polar reconstruction is the only change from OLM-SGD.

OLM-SGD (Class-Mean Gradients). This variant retains the OLM reconstruction but replaces each individual-example gradient with one class-mean gradient computed from all K support examples of that class. We select up to $B_{\mathrm{cls}} = \min\{64, C\}$ classes per update. If M_t^{cm} classes are selected at update t, their class-mean gradients are computed from KM_t^{cm} support examples in total. Its learning rate, rank, and standard epoch budget match Table 17.

Each epoch visits every support example. However, class-mean merging makes $\lceil C/B_{\mathrm{cls}} \rceil$ updates per epoch, whereas OLM-SGD makes $K\lceil C/B \rceil$. Thus, equal epoch counts do not imply equal update counts when $K > 1$. We do not multiply the learning rate by K or repeat class groups to compensate. This preserves the distinction established in Appendix C.

OLM-SGD (Class-Unique Mini-Batching). This is the proposed optimization technique and is denoted simply by OLM-SGD throughout the paper. We update all parameters within every Transformer residual block of both encoders. We set $r = 3$ and the tensor rank r_W follows (24).

Table 17: Common hyperparameters for OLM-SGD. Ablation variants retain the same learning rate, epoch budget, gradient rank, and adapted encoders, while their specific settings are described separately.

Setting	Value
Learning rate	10^{-4}
Epochs (E)	70
Class-unique capacity (B)	$\min\{64, C\}$
Gradient rank (r)	3
Adapted encoders	Vision and text

Learning-rate, training-horizon, and causal studies. The standard few-shot schedules above are distinct from the extended trajectory experiments. For the learning-rate and training-horizon sweeps, we use $K = 1$ and CLIP (ViT-B/16) on all 11 datasets. Each run lasts $E = 300$ epochs at one of four learning rates. We use seeds 1, 2, and 3. The optimizer-specific settings remain unchanged.

Table 18: Schedules for the extended studies.

Setting	Value
Backbone	CLIP (ViT-B/16)
Epochs (E)	300
Random seeds	1, 2, 3
Training-horizon sweep shots	$K = 1$
Training-horizon sweep learning rates	$\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$
Randomized mini-batch size	64
Class-unique mini-batch size (B)	$\min\{64, C\}$
Class-mean gradients per update (B_{cls})	$\min\{64, C\}$
OLM rank	$r = 3$
Rank-sensitivity values	$r \in \{1, 3, 5, 8\}$

E Proofs and Additional Theoretical Results

This appendix gives complete statements and proofs for the exact update in Section 3. It separates tensor geometry, empirical optimization under the implemented iterator, and generalization from a reused support set. Algebraic conclusions use exact arithmetic unless an approximation is specified. No restriction is imposed on the concatenated width.

E.1 Setup, probability spaces, and empirical assumptions

The support set is balanced, with $N = CK$, $C \geq 2$, and $K \geq 1$. The full-vocabulary loss and empirical objective are defined in Appendix A. The tensor reshape maps preserve entries, so their Frobenius energies sum to the squared Euclidean norm in $\mathbb{R}^d$. Write $\widetilde{g}_t$ for the concatenation of the retained means in equation 8. Define

$$\mathcal{Q}_t = \sum_{W \in \mathcal{W}} \|s_t^W\|_2^2 = \frac{1}{B_t^2} \sum_{b=1}^{B_t} \sum_{W \in \mathcal{W}} \|\Pi_{r_W}(G_t^{(b),W})\|_F^2. \tag{49}$$

The base analysis treats deterministic support losses. Augmented observations require an augmentation-averaged objective or an additional realized-gradient discrepancy. The extended ImageNet trajectories in Appendix D therefore do not directly instantiate the deterministic fixed-support setting.

Let $\mathcal{H}_t$ contain the support and the entire randomness revealed before update t, including earlier batches and decompositions. Unrevealed shuffles are drawn according to Algorithm 1. Define $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_t]$. In particular, the epoch-start history precedes its fresh shuffles. Unless explicitly

stated otherwise, expectations condition on the support and average only training randomness. The steps $\eta_t > 0$ are deterministic, T counts updates, and $\Lambda_T = \sum_{t<T} \eta_t$. All displayed expectations are assumed finite. Set $\Delta_F = F(\Theta_0) - F_{\inf}$.

Assumption E.1 (Smooth empirical objective)**.** *The function F has an L_F-Lipschitz gradient and a lower bound $F_{\inf} > -\infty$. Here, $L_F > 0$ and $F_{\inf}$ are scalars. Smoothness is required on a region containing every update segment under consideration.*

Assumption E.2 (Conditional empirical alignment)**.** *There are scalars $\alpha_{\mathrm{olm}} > 0$ and $\beta_{\mathrm{olm}} \geq 0$ such that, almost surely,*

$$\langle \nabla F(\Theta_t), \mathbb{E}_t[\widehat{g}_t] \rangle \geq \alpha_{\mathrm{olm}} \|\nabla F(\Theta_t)\|_2^2 - \beta_{\mathrm{olm}}. \tag{50}$$

Assumption E.3 (Conditional update moment)**.** *There are scalars $M_{\mathrm{olm}}, \sigma_{\mathrm{olm}} \geq 0$ such that, almost surely,*

$$\mathbb{E}_t[\|\widehat{g}_t\|_2^2] \leq M_{\mathrm{olm}} \|\nabla F(\Theta_t)\|_2^2 + \sigma_{\mathrm{olm}}^2. \tag{51}$$

These conditional assumptions allow dependent reshuffling. Sections E.4 and E.10 give explicit conditions and an epoch-wise alternative derived from the actual update.

E.2 OLM-SPECIFIC GEOMETRY

Lemma 4.1 (restated). *Assume exact singular value and polar decompositions. The following bounds hold for wide or rank-deficient factors:*

$$\langle \widetilde{g}_t, \widehat{g}_t \rangle \geq \sum_{W \in \mathcal{W}} \frac{\|\widetilde{G}_t^W\|_F^2}{\|U_t^W\|_2 \|V_t^W\|_2} \geq B_t^{-1} \|\widetilde{g}_t\|_2^2, \tag{13}$$

$$\|\widehat{g}_t\|_2^2 \leq \mathcal{Q}_t \leq B_t^{-2} \sum_{b=1}^{B_t} \|g_t^{(b)}\|_2^2. \tag{14}$$

The equality $\|\widehat{g}_t\|_2^2 = \mathcal{Q}_t$ holds when both polar factors have orthonormal columns for every tensor.

Proof. Fix a tensor W and an iteration t. Suppress only these indices, so $U \in \mathbb{R}^{m_W \times R}$, $V \in \mathbb{R}^{n_W \times R}$, and $s \in \mathbb{R}^R$ are the factors from equations 6–7, with $R = r_W B_t$. Their SVD-selected polar factors are U_{pol} and V_{pol}. Define the positive-semidefinite matrices

$$H_U = (U^\top U)^{1/2}, \qquad H_V = (V^\top V)^{1/2} \quad \text{in } \mathbb{R}^{R \times R}. \tag{52}$$

Reduced SVD gives

$$U = U_{\mathrm{pol}} H_U, \quad U_{\mathrm{pol}}^\top U = H_U, \qquad V = V_{\mathrm{pol}} H_V, \quad V^\top V_{\mathrm{pol}} = H_V. \tag{53}$$

These identities remain valid for wide matrices and zero singular values. They do not require an inverse of either Gram matrix.

Using $\widetilde{G} = U \operatorname{diag}(s) V^\top$ and $\widehat{G} = U_{\mathrm{pol}} \operatorname{diag}(s) V_{\mathrm{pol}}^\top$,

$$\begin{aligned} \langle \widetilde{G}, \widehat{G} \rangle_F &= \operatorname{tr}(\operatorname{diag}(s) H_U \operatorname{diag}(s) H_V) \\ &= \left\| H_U^{1/2} \operatorname{diag}(s) H_V^{1/2} \right\|_F^2 \geq 0. \end{aligned} \tag{54}$$

Here, tr denotes matrix trace. Both polar factors have spectral norm one. Consequently,

$$\begin{aligned} \|\widetilde{G}\|_F^2 &= \|U_{\mathrm{pol}} H_U \operatorname{diag}(s) H_V V_{\mathrm{pol}}^\top\|_F^2 \\ &\leq \|H_U\|_2 \|H_V\|_2 \left\| H_U^{1/2} \operatorname{diag}(s) H_V^{1/2} \right\|_F^2 \\ &= \|U\|_2 \|V\|_2 \langle \widetilde{G}, \widehat{G} \rangle_F. \end{aligned} \tag{55}$$

Each individual retained factor has orthonormal columns. Thus $UU^\top = \sum_{b=1}^{B_t} U^{(b)} U^{(b)\top} \preceq B_t I_{m_W}$ and, similarly, $VV^\top \preceq B_t I_{n_W}$. The notation $\preceq$ denotes positive-semidefinite order. Hence $1 \leq \|U\|_2, \|V\|_2 \leq \sqrt{B_t}$, and

$$\langle \widetilde{G}_t^W, \widehat{G}_t^W \rangle_F \geq \frac{\|\widetilde{G}_t^W\|_F^2}{\|U_t^W\|_2 \|V_t^W\|_2} \geq \frac{\|\widetilde{G}_t^W\|_F^2}{B_t}. \tag{56}$$

Summing over W proves equation 13. In particular, the inner product is strictly positive whenever the retained mean is nonzero.

For the norm bound,

$$\|\widehat{G}_t^W\|_F \le \|U_{t,\mathrm{pol}}^W\|_2 \|\operatorname{diag}(s_t^W)\|_F \|V_{t,\mathrm{pol}}^W\|_2 = \|s_t^W\|_2. \tag{57}$$

Exact truncation does not increase Frobenius energy. Sum the squared tensor inequalities and use that the fixed reshape maps preserve entries. This proves equation 14. Equality with $\mathcal{Q}_t$ holds if both polar factors have orthonormal columns for every tensor. It is not generally valid for wide factors. □

Interpretation. Polar reconstruction preserves alignment with the retained mean even when its distortion from that mean is large. Small polar distortion is therefore sufficient for some analyses but is not necessary for this alignment result. The result does not say that each original example receives a descent direction. It also does not imply that the update vanishes whenever the retained mean is zero.

Zero and single-example cases. If all individual gradients vanish then $s_t^W = 0$ for every tensor and the update vanishes, irrespective of the SVD completion. For $B_t = 1$, the retained factors already have orthonormal columns, so reconstruction equals the truncated gradient. If its retained rank is full, this is SGD. By contrast, a zero retained mean alone need not give a zero update. For $G^{(1)} = e_1 e_1^\top$ and $G^{(2)} = -e_1 e_1^\top$ in $\mathbb{R}^{2\times 2}$ with $r = 1$, the retained mean is zero but the two square completed polar factors yield $\|\widehat{G}\|_F^2 = 1/2$. This follows from $s = (1/2, 1/2)^\top$ and the tall-case energy identity, independently of the chosen completion. Thus an additive moment term can be necessary even without truncation error.

Approximate decompositions. If a numerical routine returns orthonormal retained factors and nonnegative singular coefficients, equations 54–56 remain exact for the reconstructed retained matrices supplied by that routine. The identity with the best-rank map Π_{r_W} must then be replaced by those actual reconstructions. The last inequality in equation 14 additionally requires their energies not to exceed the input energies. Otherwise their numerical excess must be retained. For any approximation, the empirical theorem applies to the actual update if its assumptions hold. No exact best-rank or numerical-equivalence claim follows merely from calling a randomized SVD routine.

E.3 Exact Energy Accounting and Factor Geometry

Proposition E.4 (Exact rectangular energy accounting)**.** *Fix a tensor and an update. Write* $\mathsf{S} = \operatorname{diag}(s)$*,* $\widetilde{G} = U\mathsf{S}V^\top$*, and* $\widehat{G} = U_{\mathrm{pol}}\mathsf{S}V_{\mathrm{pol}}^\top$*, using the reduced-SVD polar convention in equation 9. Define the* $R \times R$ *orthogonal projectors*

$$\mathsf{P}_U = U_{\mathrm{pol}}^\top U_{\mathrm{pol}}, \qquad \mathsf{P}_V = V_{\mathrm{pol}}^\top V_{\mathrm{pol}}.$$

These projectors include the SVD-selected null-space completions. Then

$$\|\widehat{G}\|_F^2 = \|\mathsf{P}_U\mathsf{S}\mathsf{P}_V\|_F^2 = \|s\|_2^2 - \|\mathsf{S} - \mathsf{P}_U\mathsf{S}\mathsf{P}_V\|_F^2. \tag{58}$$

In particular, equality with $\|s\|_2^2$ *holds exactly when* $\mathsf{S} = \mathsf{P}_U\mathsf{S}\mathsf{P}_V$*. It holds whenever both concatenated factors are tall, including square and rank-deficient factors under the stated completion convention.*

Proof. The reduced SVD gives $\mathsf{P}_U^2 = \mathsf{P}_U = \mathsf{P}_U^\top$ and the corresponding identities for P_V. Cyclic invariance of trace therefore gives

$$\|\widehat{G}\|_F^2 = \operatorname{tr}(\mathsf{S}\mathsf{P}_U\mathsf{S}\mathsf{P}_V) = \|\mathsf{P}_U\mathsf{S}\mathsf{P}_V\|_F^2.$$

The linear map $X \mapsto \mathsf{P}_U X \mathsf{P}_V$ is idempotent and self-adjoint for the Frobenius inner product. It is consequently an orthogonal projection on the matrix space. Pythagoras proves the second equality. For a tall input, its completed polar factor has orthonormal columns, so its projector equals I_R. □

For exact truncation, this identity separates both sources of energy reduction. Restoring indices gives the exact equality

$$\begin{aligned}\|\widehat{g}_t\|_2^2 = \frac{1}{B_t^2}\sum_{b=1}^{B_t}\|g_t^{(b)}\|_2^2 - \frac{1}{B_t^2}\sum_{b=1}^{B_t}\sum_{W}\|G_t^{(b),W} - \Pi_{r_W}(G_t^{(b),W})\|_F^2 \\ - \sum_{W}\|\mathsf{S}_t^W - \mathsf{P}_{U,t}^W \mathsf{S}_t^W \mathsf{P}_{V,t}^W\|_F^2,\end{aligned} \tag{59}$$

where $\mathsf{S}_t^W = \mathrm{diag}(s_t^W)$. The first subtraction is the orthogonal SVD residual. The second is the rectangular reconstruction gap. Neither comparison refers to the squared norm of the arithmetic mean gradient.

E.4 Quantitative Empirical Sufficient Conditions

Define the conditional empirical bias by

$$\mathfrak{b}_t = \|\mathbb{E}_t[\widehat{g}_t] - \nabla F(\Theta_t)\|_2. \tag{60}$$

The following nonnegative scalar distortions preserve the definitions used in the empirical analysis:

$$\varepsilon_{t,\mathrm{tail}} = \left[\sum_{W\in\mathcal{W}}\left(\frac{1}{B_t}\sum_{b=1}^{B_t}\|G_t^{(b),W} - \Pi_{r_W}(G_t^{(b),W})\|_F\right)^2\right]^{1/2}, \tag{61}$$

$$\varepsilon_{t,\mathrm{polar}} = \left[\sum_{W\in\mathcal{W}}\|\widehat{G}_t^W - \widetilde{G}_t^W\|_F^2\right]^{1/2}, \tag{62}$$

$$\varepsilon_{t,\mathrm{sample}} = \|\mathbb{E}_t[g_t] - \nabla F(\Theta_t)\|_2. \tag{63}$$

Here $\xi_t = g_t - \nabla F(\Theta_t) \in \mathbb{R}^d$ is the realized batch error. The first quantity measures discarded matrix components. The second measures the change induced by polar reconstruction. The third is conditional sampling bias, not sampling variance. All matrix and vector dimensions are inherited from Section 3.

Lemma E.5 (Empirical bias decomposition)**.** *The actual conditional bias in equation 60 satisfies*

$$\mathfrak{b}_t \le \varepsilon_{t,\mathrm{sample}} + \mathbb{E}_t[\varepsilon_{t,\mathrm{tail}} + \varepsilon_{t,\mathrm{polar}}]. \tag{64}$$

Suppose this upper bound is at most

$$\rho_{\mathrm{olm}}\|\nabla F(\Theta_t)\|_2 + \delta_{\mathrm{olm}}, \tag{65}$$

for scalars $0 \le \rho_{\mathrm{olm}} < 1$ and $\delta_{\mathrm{olm}} \ge 0$. When $\delta_{\mathrm{olm}} = 0$, Assumption E.2 holds with $\alpha_{\mathrm{olm}} = 1-\rho_{\mathrm{olm}}$ and $\beta_{\mathrm{olm}} = 0$. In general it holds with

$$\alpha_{\mathrm{olm}} = \frac{1-\rho_{\mathrm{olm}}}{2}, \qquad \beta_{\mathrm{olm}} = \frac{\delta_{\mathrm{olm}}^2}{2(1-\rho_{\mathrm{olm}})}. \tag{66}$$

Proof. For each tensor, the triangle inequality for the average discarded matrices yields $\|\widetilde{g}_t - g_t\|_2 \le \varepsilon_{t,\mathrm{tail}}$. By construction, $\|\widehat{g}_t - \widetilde{g}_t\|_2 = \varepsilon_{t,\mathrm{polar}}$. Insert g_t between $\widehat{g}_t$ and the full empirical gradient, then apply conditional Jensen and the triangle inequality. This gives equation 64.

For the second claim,

$$\begin{aligned}\langle \nabla F(\Theta_t), \mathbb{E}_t[\widehat{g}_t]\rangle &\ge \|\nabla F(\Theta_t)\|_2^2 - \|\nabla F(\Theta_t)\|_2 \mathfrak{b}_t \\ &\ge (1-\rho_{\mathrm{olm}})\|\nabla F(\Theta_t)\|_2^2 - \delta_{\mathrm{olm}}\|\nabla F(\Theta_t)\|_2.\end{aligned} \tag{67}$$

The zero-additive-error case follows directly. Otherwise, Young's inequality bounds the last product by $(1-\rho_{\mathrm{olm}})\|\nabla F(\Theta_t)\|_2^2/2 + \delta_{\mathrm{olm}}^2/[2(1-\rho_{\mathrm{olm}})]$. □

Lemma E.6 (Component growth implies update-moment control)**.** *Suppose $B_t \geq B_{\min} \geq 1$ and, almost surely,*

$$\frac{1}{B_t}\sum_{b=1}^{B_t}\mathbb{E}_t[\|g_t^{(b)}\|_2^2] \leq A_{\text{grad}}\|\nabla F(\Theta_t)\|_2^2 + \nu_{\text{grad}}^2, \tag{68}$$

where $B_{\min}$ is an integer and $A_{\text{grad}}, \nu_{\text{grad}} \geq 0$ are scalars. Then Assumption E.3 holds with

$$M_{\text{olm}} = A_{\text{grad}}/B_{\min}, \qquad \sigma_{\text{olm}}^2 = \nu_{\text{grad}}^2/B_{\min}. \tag{69}$$

Proof. Take conditional expectation in equation 14 and apply equation 68. Batch sizes in the balanced iterator are deterministic at a given batch position, including the final chunk. The bound by $B_{\min}$ follows immediately. □

An alignment certificate that does not penalize polar distortion. Let $\varepsilon_{t,\text{target}} \geq 0$ be any pathwise upper bound on $\|\widetilde{g}_t - \nabla F(\Theta_t)\|_2$. For example, $\varepsilon_{t,\text{target}} = \varepsilon_{t,\text{tail}} + \|\xi_t\|_2$ is valid. Lemma 4.1 gives

$$\langle \nabla F(\Theta_t), \widehat{g}_t\rangle \geq \frac{1}{B_t}\big(\|\nabla F(\Theta_t)\|_2 - \varepsilon_{t,\text{target}}\big)_+^2 - \varepsilon_{t,\text{target}}\sqrt{\mathcal{Q}_t}, \tag{70}$$

where $(u)_+ = \max(u, 0)$ for a real scalar u. This follows by adding and subtracting $\widetilde{g}_t$, applying Cauchy–Schwarz, and using the reverse triangle inequality. Consequently, if for every realized update $\varepsilon_{t,\text{target}} \leq \rho_{\text{ret}}\|\nabla F(\Theta_t)\|_2$ and $\sqrt{\mathcal{Q}_t} \leq \kappa_{\text{ret}}\|\nabla F(\Theta_t)\|_2$, with scalar constants $0 \leq \rho_{\text{ret}} < 1$ and $\kappa_{\text{ret}} \geq 0$, then

$$\alpha_{\text{olm}} = \frac{(1-\rho_{\text{ret}})^2}{B} - \rho_{\text{ret}}\kappa_{\text{ret}} > 0, \quad \beta_{\text{olm}} = 0, \quad M_{\text{olm}} = \kappa_{\text{ret}}^2, \quad \sigma_{\text{olm}} = 0 \tag{71}$$

are sufficient. These are strong, explicit conditions. They are not asserted to hold throughout CLIP adaptation. They show how retained information can establish alignment without requiring the polar factors to remain close to their inputs.

E.5 Polar distortion under small factor overlap

Proposition E.7 (A sharper overlap perturbation bound)**.** *Suppose $U \in \mathbb{R}^{m\times R}$ and $V \in \mathbb{R}^{n\times R}$ satisfy $R \leq \min(m, n)$ and*

$$\chi_U = \|U^\top U - I_R\|_2 < 1, \qquad \chi_V = \|V^\top V - I_R\|_2 < 1.$$

For $0 \leq x < 1$, set $\psi(x) = 1 - \sqrt{1-x}$. Then

$$\begin{aligned}\|\widehat{G} - \widetilde{G}\|_F \leq \min\{&\psi(\chi_U) + \sqrt{1+\chi_U}\,\psi(\chi_V),\\ &\psi(\chi_V) + \sqrt{1+\chi_V}\,\psi(\chi_U)\}\,\|s\|_2.\end{aligned} \tag{72}$$

Proof. All singular values of U lie in $[\sqrt{1-\chi_U}, \sqrt{1+\chi_U}]$. In its reduced SVD, the polar operation replaces these singular values by one. Hence

$$\|U_{\text{pol}} - U\|_2 = \max_j |1 - \sigma_j(U)| \leq 1 - \sqrt{1-\chi_U},$$

where the last inequality uses $\sqrt{1+x} - 1 \leq 1 - \sqrt{1-x}$ for $0 \leq x < 1$. Also $\|U\|_2 \leq \sqrt{1+\chi_U}$. Apply the same argument to V. Use the two decompositions

$$\widehat{G} - \widetilde{G} = (U_{\text{pol}} - U)\,\text{diag}(s)V_{\text{pol}}^\top + U\,\text{diag}(s)(V_{\text{pol}} - V)^\top$$

and

$$\widehat{G} - \widetilde{G} = U_{\text{pol}}\,\text{diag}(s)(V_{\text{pol}} - V)^\top + (U_{\text{pol}} - U)\,\text{diag}(s)V^\top.$$

The triangle inequality and $\|AXB\|_F \leq \|A\|_2\|X\|_F\|B\|_2$ give the two bounds in equation 72. □

E.6 Convergence from Quantified Retained-Gradient Errors

Theorem E.8 (Stationarity with explicit geometric residuals)**.** *Consider the implemented class-unique iterator and exact OLM decompositions. Suppose F is bounded below by $F_{\inf}$ and has an L_F-Lipschitz gradient on all update segments. Let B be the batch capacity and define*

$$\varepsilon_{t,\mathrm{target}} = \|\widetilde{g}_t - \nabla F(\Theta_t)\|_2, \qquad \Lambda_T = \sum_{t=0}^{T-1} \eta_t, \qquad \Delta_F = F(\Theta_0) - F_{\inf}.$$

For deterministic positive step sizes and finite displayed moments,

$$\begin{aligned}\frac{1}{\Lambda_T}\sum_{t<T}\eta_t\mathbb{E}[\|\nabla F(\Theta_t)\|_2^2] \le \frac{2B\Delta_F}{\Lambda_T} + \frac{2}{\Lambda_T}\sum_{t<T}\eta_t\mathbb{E}[\varepsilon_{t,\mathrm{target}}^2] \\ + \frac{2B}{\Lambda_T}\sum_{t<T}\eta_t\mathbb{E}[\varepsilon_{t,\mathrm{target}}\sqrt{\mathcal{Q}_t}] + \frac{BL_F}{\Lambda_T}\sum_{t<T}\eta_t^2\mathbb{E}[\mathcal{Q}_t].\end{aligned} \tag{73}$$

This statement does not require conditional unbiasedness or independent batches. Its residual is bounded by $\varepsilon_{t,\mathrm{target}} \le \varepsilon_{t,\mathrm{tail}} + \|g_t - \nabla F(\Theta_t)\|_2$.

Proof. Write $f_t = \nabla F(\Theta_t)$. Lemma 4.1 and Cauchy–Schwarz give

$$\langle f_t, \widehat{g}_t\rangle \ge \frac{1}{B}\|\widetilde{g}_t\|_2^2 - \varepsilon_{t,\mathrm{target}}\sqrt{\mathcal{Q}_t} \ge \frac{1}{2B}\|f_t\|_2^2 - \frac{1}{B}\varepsilon_{t,\mathrm{target}}^2 - \varepsilon_{t,\mathrm{target}}\sqrt{\mathcal{Q}_t}.$$

The last step follows from $\|f_t\|_2^2 \le 2\|\widetilde{g}_t\|_2^2 + 2\varepsilon_{t,\mathrm{target}}^2$. Apply the descent lemma and $\|\widehat{g}_t\|_2^2 \le \mathcal{Q}_t$. Take expectations, sum, and use $\mathbb{E}[F(\Theta_T)] \ge F_{\inf}$. The final residual inequality follows by inserting g_t and using the tensor-wise truncation residual bound. □

Proposition E.9 (Conditional geometric sufficient conditions)**.** *Let $\mathcal{H}_t$ contain the fixed support and the complete revealed training history before update t. Set $f_t = \nabla F(\Theta_t)$ and $\varepsilon_{t,\mathrm{target}} = \|\widetilde{g}_t - f_t\|_2$. Suppose deterministic constants $0 \le \rho_{\mathrm{ret}} < 1$, $\kappa_{\mathrm{ret}}, \delta_{\mathrm{ret}}, \nu_{\mathrm{ret}} \ge 0$ satisfy, almost surely,*

$$\sqrt{\mathbb{E}_t[\varepsilon_{t,\mathrm{target}}^2]} \le \rho_{\mathrm{ret}}\|f_t\|_2 + \delta_{\mathrm{ret}}, \qquad \sqrt{\mathbb{E}_t[\mathcal{Q}_t]} \le \kappa_{\mathrm{ret}}\|f_t\|_2 + \nu_{\mathrm{ret}}. \tag{74}$$

Define

$$a_{\mathrm{ret}} = \frac{(1-\rho_{\mathrm{ret}})^2}{B} - \rho_{\mathrm{ret}}\kappa_{\mathrm{ret}}, \qquad c_{\mathrm{ret}} = \frac{2(1-\rho_{\mathrm{ret}})\delta_{\mathrm{ret}}}{B} + \rho_{\mathrm{ret}}\nu_{\mathrm{ret}} + \kappa_{\mathrm{ret}}\delta_{\mathrm{ret}}.$$

If $a_{\mathrm{ret}} > 0$, the conditional alignment assumption holds with

$$\alpha_{\mathrm{olm}} = a_{\mathrm{ret}}/2, \qquad \beta_{\mathrm{olm}} = c_{\mathrm{ret}}^2/(2a_{\mathrm{ret}}) + \delta_{\mathrm{ret}}\nu_{\mathrm{ret}}.$$

The update-moment assumption holds with $M_{\mathrm{olm}} = 2\kappa_{\mathrm{ret}}^2$ and $\sigma_{\mathrm{olm}}^2 = 2\nu_{\mathrm{ret}}^2$. If $\delta_{\mathrm{ret}} = \nu_{\mathrm{ret}} = 0$, the sharper constants $\alpha_{\mathrm{olm}} = a_{\mathrm{ret}}$, $\beta_{\mathrm{olm}} = 0$, $M_{\mathrm{olm}} = \kappa_{\mathrm{ret}}^2$, and $\sigma_{\mathrm{olm}} = 0$ hold.

Proof. The reverse triangle inequality in conditional L^2 gives

$$\sqrt{\mathbb{E}_t[\|\widetilde{g}_t\|_2^2]} \ge \big(\|f_t\|_2 - \sqrt{\mathbb{E}_t[\varepsilon_{t,\mathrm{target}}^2]}\big)_+.$$

The geometric lemma and conditional Cauchy–Schwarz therefore imply

$$\langle f_t, \mathbb{E}_t[\widehat{g}_t]\rangle \ge \frac{((1-\rho_{\mathrm{ret}})\|f_t\|_2 - \delta_{\mathrm{ret}})_+^2}{B} - (\rho_{\mathrm{ret}}\|f_t\|_2 + \delta_{\mathrm{ret}})(\kappa_{\mathrm{ret}}\|f_t\|_2 + \nu_{\mathrm{ret}}).$$

For nonnegative x, δ_{ret}, $(x-\delta_{\mathrm{ret}})_+^2 \ge x^2 - 2x\delta_{\mathrm{ret}}$. Expand to obtain the lower bound $a_{\mathrm{ret}}\|f_t\|_2^2 - c_{\mathrm{ret}}\|f_t\|_2 - \delta_{\mathrm{ret}}\nu_{\mathrm{ret}}$. Young's inequality yields the stated alignment constants. The energy bound and $(a+b)^2 \le 2a^2 + 2b^2$ give the moment constants. When both additive errors vanish, neither use of these inequalities is necessary, giving the sharper constants. □

E.7 Conditional empirical convergence

Theorem E.10 (Conditional empirical stationarity)**.** *Under Assumptions E.1–E.3, suppose*

$$0 < \eta_t \le \frac{\alpha_{\mathrm{olm}}}{L_F M_{\mathrm{olm}}}, \tag{75}$$

with no upper restriction from this expression if $M_{\mathrm{olm}} = 0$. Set $\Delta_F = F(\Theta_0) - F_{\mathrm{inf}} \ge 0$. Then

$$\frac{1}{\Lambda_T}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[\|\nabla F(\Theta_t)\|_2^2] \le \frac{2\Delta_F}{\alpha_{\mathrm{olm}}\Lambda_T} + \frac{2\beta_{\mathrm{olm}}}{\alpha_{\mathrm{olm}}} + \frac{L_F\sigma_{\mathrm{olm}}^2}{\alpha_{\mathrm{olm}}\Lambda_T}\sum_{t=0}^{T-1}\eta_t^2. \tag{76}$$

Proof. Smoothness along the update segment gives

$$F(\Theta_{t+1}) \le F(\Theta_t) - \eta_t\langle\nabla F(\Theta_t), \widehat{g}_t\rangle + \frac{L_F\eta_t^2}{2}\|\widehat{g}_t\|_2^2. \tag{77}$$

Conditional expectation and the two update assumptions imply

$$\begin{aligned}\mathbb{E}_t[F(\Theta_{t+1})] \le F(\Theta_t) - \eta_t\left(\alpha_{\mathrm{olm}} - \frac{L_F M_{\mathrm{olm}}\eta_t}{2}\right)\|\nabla F(\Theta_t)\|_2^2 \\ + \eta_t\beta_{\mathrm{olm}} + \frac{L_F\eta_t^2\sigma_{\mathrm{olm}}^2}{2}.\end{aligned} \tag{78}$$

Equation 75 makes the coefficient in parentheses at least $\alpha_{\mathrm{olm}}/2$. Taking expectation over training randomness, telescoping, and using the lower bound gives

$$\frac{\alpha_{\mathrm{olm}}}{2}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[\|\nabla F(\Theta_t)\|_2^2] \le \Delta_F + \beta_{\mathrm{olm}}\Lambda_T + \frac{L_F\sigma_{\mathrm{olm}}^2}{2}\sum_{t=0}^{T-1}\eta_t^2. \tag{79}$$

Division proves equation 76. For a constant $\eta_t = \eta$,

$$\min_{0\le t<T}\mathbb{E}[\|\nabla F(\Theta_t)\|_2^2] \le \frac{2\Delta_F}{\alpha_{\mathrm{olm}}\eta T} + \frac{2\beta_{\mathrm{olm}}}{\alpha_{\mathrm{olm}}} + \frac{L_F\eta\sigma_{\mathrm{olm}}^2}{\alpha_{\mathrm{olm}}}. \tag{80}$$

The minimum is bounded by the arithmetic average, but it is not necessarily attained at the last iterate. □

If $\beta_{\mathrm{olm}} = 0$, $\Lambda_T \to \infty$, and $\sum_{t<T}\eta_t^2/\Lambda_T \to 0$, the weighted average vanishes. Square-summable steps with a divergent sum are sufficient. This statement is not almost-sure convergence of all parameters or all gradient norms. With $\sigma_{\mathrm{olm}} = \beta_{\mathrm{olm}} = 0$, a fixed admissible step gives an $\mathcal{O}(T^{-1})$ bound.

Equivalently, sample an index I_T independently of the trajectory with $\Pr(I_T = t) = \eta_t/\Lambda_T$. The bound is on $\mathbb{E}[\|\nabla F(\Theta_{I_T})\|_2^2]$. Neither last-iterate stationarity nor parameter convergence is asserted. The $T^{-1/2}$ finite-horizon specialization assumes $\beta_{\mathrm{olm}} = 0$ and an admissible constant step $\eta = \Theta(T^{-1/2})$.

E.8 A Polyak–Łojasiewicz Specialization

Corollary E.11 (Empirical objective gap under PL)**.** *In addition to Theorem E.10, assume the scalar F_{inf} equals the objective infimum and*

$$\frac{1}{2}\|\nabla F(\Theta)\|_2^2 \ge \mu_{\mathrm{PL}}\big(F(\Theta) - F_{\mathrm{inf}}\big), \tag{81}$$

for a scalar $\mu_{\mathrm{PL}} > 0$ on the relevant region. For a constant step satisfying equation 75 and $\eta \le (\alpha_{\mathrm{olm}}\mu_{\mathrm{PL}})^{-1}$,

$$\mathbb{E}[F(\Theta_T) - F_{\mathrm{inf}}] \le (1 - \alpha_{\mathrm{olm}}\mu_{\mathrm{PL}}\eta)^T\Delta_F + \frac{\beta_{\mathrm{olm}} + L_F\eta\sigma_{\mathrm{olm}}^2/2}{\alpha_{\mathrm{olm}}\mu_{\mathrm{PL}}}. \tag{82}$$

Proof. Insert equation 81 into equation 78, using the coefficient lower bound $\alpha_{\mathrm{olm}}/2$. The expected gap obeys a scalar affine recursion with multiplier $1 - \alpha_{\mathrm{olm}}\mu_{\mathrm{PL}}\eta \in [0, 1)$. Unroll it and bound its geometric sum by $(\alpha_{\mathrm{olm}}\mu_{\mathrm{PL}}\eta)^{-1}$. This proves the result. The PL property is an additional analytical regime, not an established global property of CLIP fine-tuning. □

E.9 Exact Fixed-Parameter Sampling Variances

Reindex the existing support gradients by class and shot: $g_{c,k}(\Theta) \in \mathbb{R}^d$ is $\nabla_\Theta \ell_i(\Theta)$ for the corresponding support example. This analytical reindexing does not average the shots before OLM. Define the empirical class and support means,

$$\mu_c(\Theta) = \frac{1}{K}\sum_{k=1}^{K} g_{c,k}(\Theta), \qquad \mu(\Theta) = \frac{1}{C}\sum_{c=1}^{C} \mu_c(\Theta) = \nabla F(\Theta), \tag{83}$$

both in $\mathbb{R}^d$, and the scalar dispersions

$$\Gamma_{\text{within}}(\Theta) = \frac{1}{CK}\sum_{c=1}^{C}\sum_{k=1}^{K} \|g_{c,k}(\Theta) - \mu_c(\Theta)\|_2^2, \tag{84}$$

$$\Gamma_{\text{between}}(\Theta) = \frac{1}{C}\sum_{c=1}^{C} \|\mu_c(\Theta) - \mu(\Theta)\|_2^2. \tag{85}$$

Proposition E.12 (Variance of a class-unique batch)**.** *Fix the support set and Θ independently of a new batch draw. Select $B \le C$ distinct classes uniformly, then one shot independently and uniformly within each selected class. Let $g_{\text{class}}(\Theta) \in \mathbb{R}^d$ denote the ordinary mean of these gradients. Then*

$$\mathbb{E}[g_{\text{class}}(\Theta)] = \mu(\Theta), \qquad \mathbb{E}[\|g_{\text{class}}(\Theta) - \mu(\Theta)\|_2^2] = \frac{\Gamma_{\text{within}}(\Theta)}{B} + \frac{C-B}{B(C-1)}\Gamma_{\text{between}}(\Theta). \tag{86}$$

The ordinary mean from B independent uniform support draws has variance

$$\frac{\Gamma_{\text{within}}(\Theta) + \Gamma_{\text{between}}(\Theta)}{B}, \tag{87}$$

whereas B uniformly selected distinct support examples have variance

$$\frac{N-B}{B(N-1)}\big(\Gamma_{\text{within}}(\Theta) + \Gamma_{\text{between}}(\Theta)\big). \tag{88}$$

Class-unique variance minus the last variance is

$$\frac{B-1}{B(N-1)}\left(\Gamma_{\text{within}}(\Theta) - \frac{N-C}{C-1}\Gamma_{\text{between}}(\Theta)\right). \tag{89}$$

Proof. For a selected class c, write the selected gradient as its empirical mean plus a centered within-class residual. Conditioned on the chosen class subset, those residuals are independent and have zero mean. Averaging over subsets makes their contribution Γ_{within}/B.

For the class means, $\sum_c(\mu_c - \mu) = 0$. The inclusion probability of one class is B/C, and the joint inclusion probability of two distinct classes is $B(B-1)/[C(C-1)]$. Expanding the squared mean and using $\sum_{c\neq c'}\langle \mu_c - \mu, \mu_{c'} - \mu\rangle = -C\Gamma_{\text{between}}$ gives $(C-B)\Gamma_{\text{between}}/[B(C-1)]$. The cross term with within-class residuals has zero expectation. This proves equation 86. Independent draws give equation 87. Repeating the inclusion-probability calculation for the N individual support examples gives equation 88. Subtraction and $N = CK$ yield equation 89. □

At $K = 1$, $\Gamma_{\text{within}} = 0$ and $N = C$, so the two without-replacement batch laws agree. This does not imply that the one-shot support estimates the population without error. At $B = C$, class selection contributes no variance, but choosing one of K shots per class still contributes Γ_{within}/C. Smaller final chunks obey the same fixed-parameter formula with their actual size B_t. None of these statements transfers an input variance ordering through a nonlinear merge.

Fresh-data supervised training. Using the class-conditional distributions and population quantities defined in equations 105 and 113, consider a separate sampling regime in which a new image is available from every selected class at each step. This is not the finite-support epoch iterator. Define $\mu^{\text{pop}}(\Theta) = C^{-1}\sum_c \mu_c^{\text{pop}}(\Theta) \in \mathbb{R}^d$ and $\Gamma_{\text{between}}^{\text{pop}}(\Theta) = C^{-1}\sum_c \|\mu_c^{\text{pop}}(\Theta) - \mu^{\text{pop}}(\Theta)\|_2^2$. The ordinary class-unique mean is conditionally unbiased for $\nabla\mathcal{R}(\Theta_t)$ and has variance

$$\frac{\Gamma_{\text{within}}^{\text{pop}}(\Theta_t)}{B} + \frac{C-B}{B(C-1)}\Gamma_{\text{between}}^{\text{pop}}(\Theta_t). \tag{90}$$

The proof is the same decomposition with population means and fresh independent class samples. There is no persistent support-draw discrepancy in this oracle regime. OLM merging can nevertheless remain biased. The proof of Theorem E.10 applies to $\mathcal{R}$ after imposing the corresponding population alignment, smoothness, and moment assumptions. Those conditions do not follow from fresh sampling alone. For ordinary supervised training on a finite dataset, the finite-support analysis remains the appropriate one, even when K is large.

E.10 An Epoch-Wise Certificate for the Implemented Reshuffling

Let $J = K\lceil C/B \rceil$ be the integer number of updates per epoch. Epoch $e \in \{1, \dots, E\}$ begins at iteration $t_e = (e-1)J$ and uses the set of iteration indices $\mathcal{T}_e = \{t_e, \dots, t_e + J - 1\}$. Define the frozen-parameter remainder

$$\omega_e = \frac{1}{J} \sum_{t \in \mathcal{T}_e} \frac{1}{B_t} \sum_{b=1}^{B_t} \nabla_\Theta \ell_{i_{t,b}}(\Theta_{t_e}) - \nabla F(\Theta_{t_e}) \in \mathbb{R}^d. \tag{91}$$

This measures unequal weighting caused by averaging batch means rather than examples. It is not a change to the optimizer. If B divides C, every batch has size B, every support example appears once, and $\omega_e = 0$ for every permutation. Otherwise the remainder need not vanish for one realized epoch. With fresh independent epoch shuffles, $\mathbb{E}[\omega_e \mid \mathcal{H}_{t_e}] = 0$ when that history precedes the epoch shuffle: each fixed batch position has a uniform support marginal at the frozen parameter value.

Proposition E.13 (Reshuffling without an unbiased-step substitution)**.** *Suppose each empirical per-example loss has an L_ℓ-Lipschitz gradient on every segment joining an iterate to its epoch start, and use exact OLM decompositions. Then*

$$\begin{aligned} \left\| \frac{1}{J} \sum_{t \in \mathcal{T}_e} \widehat{g}_t - \nabla F(\Theta_{t_e}) \right\|_2 &\le \frac{1}{J} \sum_{t \in \mathcal{T}_e} (\varepsilon_{t,\mathrm{tail}} + \varepsilon_{t,\mathrm{polar}}) + \|\omega_e\|_2 \\ &\quad + \frac{L_\ell}{J} \sum_{t \in \mathcal{T}_e} \sum_{t'=t_e}^{t-1} \eta_{t'} \sqrt{\mathcal{Q}_{t'}}. \end{aligned} \tag{92}$$

An empty inner sum is zero. If the learning rate is constant within the epoch, this average direction is exactly the direction of its net parameter displacement, scaled by $J\eta_{t_e}$.

Proof. Insert the raw gradient g_t into each summand. Its difference from $\widehat{g}_t$ is bounded by the two tensor distortions. For the same realized batch, compare its mean gradient at Θ_t and at Θ_{t_e}. The difference is at most $L_\ell \|\Theta_t - \Theta_{t_e}\|_2$. The average frozen batch gradient equals $\nabla F(\Theta_{t_e}) + \omega_e$ by (91). Finally, $\|\Theta_t - \Theta_{t_e}\|_2 \le \sum_{t'=t_e}^{t-1} \eta_{t'} \sqrt{\mathcal{Q}_{t'}}$ by Lemma 4.1. Summing proves the bound. Constant within-epoch steps give $\Theta_{t_e+J} - \Theta_{t_e} = -J\eta_{t_e}(J^{-1} \sum_{t \in \mathcal{T}_e} \widehat{g}_t)$. □

The certificate separates merge distortion, unequal final-batch weighting, and within-epoch parameter drift. It is valid for the actual batch sequence and does not require conditional unbiasedness at each iterate. The terms must still be controlled to obtain an informative epoch-level convergence.

E.11 Quantified convergence for class-unique epochs

The contribution of incomplete batches. Write $q_{\mathrm{batch}} = \lceil C/B \rceil$ and let $b_j = \min\{B, C - (j-1)B\}$, $1 \le j \le q_{\mathrm{batch}}$, be the deterministic chunk sizes within a round. Define

$$a_{\mathrm{batch}} = \frac{1}{q_{\mathrm{batch}}^2} \sum_{j=1}^{q_{\mathrm{batch}}} \frac{1}{b_j} - \frac{1}{C} \ge 0. \tag{93}$$

Nonnegativity follows from Cauchy–Schwarz and $\sum_j b_j = C$. The coefficient vanishes when all chunks have the same size, including $B \mid C$.

Proposition E.14 (Exact frozen-epoch remainder moment)**.** *Condition on the fixed support and the history $\mathcal{H}_{t_e}$ immediately before the fresh epoch shuffles. Assume the within-class permutations are independent and uniform, and the class permutations are independent and uniform across rounds, independently of those within-class permutations. For the remainder ω_e in equation 91,*

$$\mathbb{E}_e[\omega_e] = 0, \qquad \nu_e^2 := \mathbb{E}_e[\|\omega_e\|_2^2] = \frac{a_{\text{batch}}}{K}\left[\Gamma_{\text{within}}(\Theta_{t_e}) + \frac{C}{C-1}\Gamma_{\text{between}}(\Theta_{t_e})\right], \tag{94}$$

where $\mathbb{E}_e[\cdot] = \mathbb{E}[\cdot \mid \mathcal{H}_{t_e}]$ and the dispersions are those of equation 85.

Proof. All gradients in this proof are evaluated at Θ_{t_e}. First condition additionally on the within-class permutations, and denote the round-k gradient from class c by $v_{c,k}$. If class c occupies chunk j in round k, its coefficient in the average of that round's q_{batch} batch means is $w_{c,k} = 1/(q_{\text{batch}} b_j)$. Put $z_{c,k} = w_{c,k} - 1/C$ and $R_k = \sum_c z_{c,k} v_{c,k}$. Uniform class permutations give

$$\mathbb{E}[z_{c,k}] = 0, \quad \mathbb{E}[z_{c,k}^2] = \frac{a_{\text{batch}}}{C}, \quad \mathbb{E}[z_{c,k} z_{c',k}] = -\frac{a_{\text{batch}}}{C(C-1)} \quad (c \neq c'). \tag{95}$$

For the last identity, use exchangeability and $\sum_c z_{c,k} = 0$. Thus, with $\bar{v}_k = C^{-1}\sum_c v_{c,k}$,

$$\mathbb{E}[\|R_k\|_2^2] = \frac{a_{\text{batch}}}{C-1}\sum_{c=1}^{C}\|v_{c,k} - \bar{v}_k\|_2^2. \tag{96}$$

Every support example is used once. Hence the average of the K uniformly class-weighted round means is exactly $\nabla F(\Theta_{t_e})$, and $\omega_e = K^{-1}\sum_k R_k$. Conditional on the within-class permutations, the R_k are independent and centered because their class permutations are independent. Their cross moments therefore vanish, even though the selected shots across rounds are dependent.

For each fixed round, the within-class permutations select independent uniform shots across classes. Expanding the empirical class means and the centered within-class residuals gives

$$\mathbb{E}\left[\frac{1}{C}\sum_c \|v_{c,k} - \bar{v}_k\|_2^2\right] = \Gamma_{\text{between}} + \frac{C-1}{C}\Gamma_{\text{within}}. \tag{97}$$

Sum the K round variances and divide by K^2. This proves both claims. □

An empirical convergence guarantee for the actual iterator. For a learning rate η_e constant within epoch e, let $\gamma_e = J\eta_e$, $f_e = \nabla F(\Theta_{t_e})$, and $h_e = J^{-1}\sum_{t\in\mathcal{T}_e}\widehat{g}_t$. The actual endpoint update is $\Theta_{t_e+J} = \Theta_{t_e} - \gamma_e h_e$. Define the exact epoch reconstruction error

$$u_e = h_e - f_e - \omega_e \in \mathbb{R}^d, \qquad D_e = \|u_e\|_2, \qquad d_e^2 = \mathbb{E}_e[D_e^2]. \tag{98}$$

Its components are computable from the realized updates and the frozen epoch-start gradients. If each empirical per-example gradient is L_ℓ-Lipschitz on the segment joining an iterate to its epoch start, and the exact energy bound holds, Proposition E.13 gives

$$\begin{aligned} D_e &\le M_e + \frac{L_\ell}{J}\sum_{t\in\mathcal{T}_e}\sum_{s=t_e}^{t-1}\eta_e\sqrt{\mathcal{Q}_s}, \\ M_e &:= \frac{1}{J}\sum_{t\in\mathcal{T}_e}\|\widehat{g}_t - g_t\|_2 \le \frac{1}{J}\sum_{t\in\mathcal{T}_e}(\varepsilon_{t,\text{tail}} + \varepsilon_{t,\text{polar}}). \end{aligned} \tag{99}$$

This exact error allows cancellation across an epoch. Its upper certificate separates realized merging, retained energy, and parameter drift.

Theorem 4.2 (restated). *Fix the support and use fresh independent epoch shuffles. Let $F_{\text{inf}} \in \mathbb{R}$ be a finite lower bound of F. Assume that ∇F is L_F-Lipschitz on each segment joining an epoch start to its endpoint. Here, $L_F > 0$ is a scalar. Gradient differences have norm at most L_F times the corresponding parameter distance. Use a deterministic scalar learning rate $\eta_e > 0$ throughout epoch e. Define the scalar epoch step $\gamma_e = J\eta_e$ and assume $0 < \gamma_e \le L_F^{-1}$. Let $\Lambda_E^{\text{epoch}} = \sum_{e=1}^E \gamma_e > 0$ be the total epoch step. Let $\Delta_F = F(\Theta_0) - F_{\text{inf}} \ge 0$ be the initial objective gap. Both quantities are scalars. If all expectations below are finite, then*

$$\frac{\sum_{e=1}^E \gamma_e \mathbb{E}[\|f_e\|_2^2]}{\Lambda_E^{\text{epoch}}} \le \frac{2\Delta_F + \sum_{e=1}^E \gamma_e(1 + L_F\gamma_e)\mathbb{E}[D_e^2]}{\Lambda_E^{\text{epoch}}} + \frac{2L_F\sum_{e=1}^E \gamma_e^2\mathbb{E}[\nu_e^2]}{\Lambda_E^{\text{epoch}}}. \tag{18}$$

Proof. The endpoint update and the assumed smoothness on its chord give

$$\mathbb{E}_e[F(\Theta_{t_e+J})] \le F(\Theta_{t_e}) - \gamma_e\langle f_e, \mathbb{E}_e[h_e]\rangle + \frac{L_F\gamma_e^2}{2}\mathbb{E}_e[\|h_e\|_2^2]. \tag{100}$$

In this proof suppress e on $f, \gamma, \omega, u, d, \nu$. Substitute $h = f + \omega + u$ and use $\mathbb{E}_e[\omega] = 0$:

$$\begin{aligned}\mathbb{E}_e[F(\Theta_{t_e+J})] \le F(\Theta_{t_e}) &- \gamma(1 - L_F\gamma/2)\|f\|_2^2 - \gamma(1 - L_F\gamma)\langle f, \mathbb{E}_e[u]\rangle \\ &+ \frac{L_F\gamma^2}{2}\left(\nu^2 + 2\mathbb{E}_e[\langle\omega, u\rangle] + \mathbb{E}_e[\|u\|_2^2]\right).\end{aligned} \tag{101}$$

Conditional Cauchy–Schwarz gives $\|\mathbb{E}_e[u]\|_2 \le d$ and $\mathbb{E}_e[\langle\omega, u\rangle] \le \nu d$. Since $1 - L_F\gamma \ge 0$, apply $\|f\|_2 d \le (\|f\|_2^2 + d^2)/2$ and $\nu d \le (\nu^2 + d^2)/2$. Collecting coefficients yields

$$\mathbb{E}_e[F(\Theta_{t_e+J})] \le F(\Theta_{t_e}) - \frac{\gamma_e}{2}\|f_e\|_2^2 + \frac{\gamma_e}{2}(1 + L_F\gamma_e)d_e^2 + L_F\gamma_e^2\nu_e^2. \tag{102}$$

Take expectations, sum over epochs, and use $F \ge F_{\inf}$. No within-epoch conditional unbiasedness has been used. □

Corollary E.15 (Finite-time rates with controlled merge residual)**.** *In Theorem 4.2, assume exact decompositions and that every empirical per-example gradient is L_ℓ-Lipschitz on each segment joining a within-epoch iterate to its epoch start. Assume also that every observed-support example gradient has norm at most H_{grad} on the iterates. Let $B_{\min} = \min_j b_j$ and*

$$c_{\mathrm{drift}} = \frac{L_\ell H_{\mathrm{grad}}(J-1)}{2J\sqrt{B_{\min}}}, \qquad M_e = \frac{1}{J}\sum_{t\in\mathcal{T}_e}\|\widehat{g}_t - g_t\|_2. \tag{103}$$

Suppose $\mathbb{E}[M_e^2] \le m^2$ and $\mathbb{E}[\nu_e^2] \le v^2$ for all epochs. For a constant $\gamma \le L_F^{-1}$,

$$\frac{1}{E}\sum_{e=1}^{E}\mathbb{E}[\|f_e\|_2^2] \le \frac{2\Delta_F}{\gamma E} + 4m^2 + 4c_{\mathrm{drift}}^2\gamma^2 + 2L_F v^2\gamma. \tag{104}$$

More generally, empirical stationarity follows if $\Lambda_E^{\mathrm{epoch}} \to \infty$, $\sum_{e\le E}\gamma_e\mathbb{E}[M_e^2]/\Lambda_E^{\mathrm{epoch}} \to 0$, and $\sum_{e\le E}\gamma_e^2/\Lambda_E^{\mathrm{epoch}} \to 0$, under the same uniform gradient and remainder-moment bounds.

Proof. Lemma 4.1 gives $\sqrt{\mathcal{Q}_t} \le H_{\mathrm{grad}}/\sqrt{B_{\min}}$. The double sum in equation 99 has $J(J-1)/2$ terms. Therefore $D_e \le M_e + c_{\mathrm{drift}}\gamma_e$ and $\mathbb{E}[D_e^2] \le 2\mathbb{E}[M_e^2] + 2c_{\mathrm{drift}}^2\gamma_e^2$. Substitute this inequality and $1 + L_F\gamma_e \le 2$ into equation 18. For the diminishing-step statement, the bounded step sizes also make $\sum_e \gamma_e^3/\Lambda_E^{\mathrm{epoch}} \to 0$ under the stated squared-step condition. □

When $m = 0$, a horizon-dependent γ of order $E^{-1/2}$ gives an $\mathcal{O}(E^{-1/2})$ bound. If the chunks are equal, then $v = 0$ and order $E^{-1/3}$ yields $\mathcal{O}(E^{-2/3})$. These are sufficient regimes of the quantitative certificate, not assertions that the observed OLM residual vanishes. A persistent residual retains the explicit $4m^2$ neighborhood. Fresh epoch permutations center incomplete-batch weighting, but do not create fresh labeled observations.

Corollary E.16 (Absorbing a relative epoch residual)**.** *Under Theorem 4.2, suppose $\gamma_e \le \gamma_{\max} \le L_F^{-1}$ and $\mathbb{E}_e[D_e^2] \le \rho_{\mathrm{epoch}}^2\|f_e\|_2^2 + \delta_e^2$ for deterministic $0 \le \rho_{\mathrm{epoch}} < 1$ and nonnegative, history-measurable δ_e. If $c_{\mathrm{epoch}} = 1 - (1 + L_F\gamma_{\max})\rho_{\mathrm{epoch}}^2 > 0$, then the bound in equation 18 holds with D_e^2 replaced by δ_e^2 and the entire right-hand side divided by c_{epoch}. For constant $\gamma_e = \gamma$, $\mathbb{E}[\delta_e^2] \le \bar{\delta}^2$, and $\mathbb{E}[\nu_e^2] \le v^2$, it is at most*

$$\frac{1}{c_{\mathrm{epoch}}}\left(\frac{2\Delta_F}{\gamma E} + (1 + L_F\gamma)\bar{\delta}^2 + 2L_F\gamma v^2\right).$$

Proof. Substitute the assumed second-moment bound into equation 18 and move the weighted gradient term to the left. Its remaining coefficient is at least c_{epoch} at every epoch. The constant-step conclusion follows by division. □

With a zero additive residual and bounded v, an admissible horizon-dependent $\gamma = \Theta(E^{-1/2})$ therefore gives $\mathcal{O}(E^{-1/2})$ stationarity even when the merge has a nonzero relative distortion. With diminishing steps, the weighted residual and squared-step terms must vanish. These are conditions on the actual epoch reconstruction error, not assumptions of independent within-epoch batches.

E.12 Population setup and trajectory-wise guarantees

Population and support are different objects. For each class c, let $\mathcal{P}_c$ be its image distribution. Draw K independent support images from each $\mathcal{P}_c$, independently across classes. The pretrained initialization and frozen model components are independent of this support draw. Write $\ell(\Theta; x, c) = -\log p_\Theta(c \mid x)$, so its support evaluations are the existing $\ell_i(\Theta)$. Define the class-balanced population objective

$$\mathcal{R}(\Theta) = \frac{1}{C}\sum_{c=1}^{C} \mathbb{E}_{x\sim\mathcal{P}_c}[\ell(\Theta; x, c)]. \tag{105}$$

Both $\mathcal{R}$ and F are scalar-valued. Differentiation under the population expectation is assumed valid. This analyzes within-task generalization to the same classes, not transfer to unseen classes or distributions.

Define the finite-support discrepancy and the batch error,

$$e_{\sup}(\Theta) = \nabla F(\Theta) - \nabla\mathcal{R}(\Theta), \qquad \xi_t = g_t - \nabla F(\Theta_t), \tag{106}$$

which lie in $\mathbb{R}^d$. Conditional on the support, this is a fixed function of Θ, not fresh noise at each update. The vector ξ_t need not be conditionally centered under reshuffling. Exactly,

$$\widehat{g}_t - \nabla\mathcal{R}(\Theta_t) = (\widehat{g}_t - g_t) + \xi_t + e_{\sup}(\Theta_t). \tag{107}$$

The terms distinguish nonlinear merging, batch selection, and finite supervision. In particular, increasing the number of epochs does not increase CK.

Theorem E.17 (Trajectory-wise few-shot population bound)**.** *Suppose $\mathcal{R}$ has an $L_{\mathcal{R}}$-Lipschitz gradient on the update segments and is bounded below by $\mathcal{R}_{\inf}$. Set $\Delta_{\mathcal{R}} = \mathcal{R}(\Theta_0) - \mathcal{R}_{\inf} \geq 0$. For any deterministic positive learning rates and finite moments,*

$$\begin{aligned}&\frac{1}{\Lambda_T}\sum_{t=0}^{T-1}\eta_t\mathbb{E}[\|\nabla\mathcal{R}(\Theta_t)\|_2^2]\\ &\quad\leq \frac{2\Delta_{\mathcal{R}}}{\Lambda_T} + \frac{1}{\Lambda_T}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\big[(\mathfrak{b}_t + \|e_{\sup}(\Theta_t)\|_2)^2\big] + \frac{L_{\mathcal{R}}}{\Lambda_T}\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}[\mathcal{Q}_t].\end{aligned} \tag{108}$$

Proof. The support and Θ_t are $\mathcal{H}_t$-measurable. From the triangle inequality,

$$\|\mathbb{E}_t[\widehat{g}_t] - \nabla\mathcal{R}(\Theta_t)\|_2 \leq \mathfrak{b}_t + \|e_{\sup}(\Theta_t)\|_2. \tag{109}$$

Cauchy–Schwarz and Young's inequality give

$$\langle\nabla\mathcal{R}(\Theta_t), \mathbb{E}_t[\widehat{g}_t]\rangle \geq \frac{1}{2}\|\nabla\mathcal{R}(\Theta_t)\|_2^2 - \frac{1}{2}(\mathfrak{b}_t + \|e_{\sup}(\Theta_t)\|_2)^2. \tag{110}$$

Apply $L_{\mathcal{R}}$-smoothness to the actual update. Take conditional expectation and use $\mathbb{E}_t[\|\widehat{g}_t\|_2^2] \leq \mathbb{E}_t[\mathcal{Q}_t]$:

$$\begin{aligned}\mathbb{E}_t[\mathcal{R}(\Theta_{t+1})] \leq \mathcal{R}(\Theta_t) &- \frac{\eta_t}{2}\|\nabla\mathcal{R}(\Theta_t)\|_2^2\\ &+ \frac{\eta_t}{2}(\mathfrak{b}_t + \|e_{\sup}(\Theta_t)\|_2)^2 + \frac{L_{\mathcal{R}}\eta_t^2}{2}\mathbb{E}_t[\mathcal{Q}_t].\end{aligned} \tag{111}$$

Take expectation over the training trajectory, telescope, apply the population lower bound, and divide by $\Lambda_T/2$. This proves equation 108. No centering of ξ_t or independence of successive batches has been used. □

The bound remains conditional on the observed support. More data can reduce the statistical discrepancy, but repeated optimization on the same data is not additional independent supervision. A constant-step bound may retain both a bias neighborhood and a moment term. Vanishing weighted empirical optimization error alone does not eliminate the finite-shot term.

E.13 A Direction-Preserving Population Certificate

Proposition E.18 (Population progress in retained directions)**.** *Under the smoothness and lower-bound conditions of Theorem E.17,*

$$\frac{1}{\Lambda_T}\sum_{t=0}^{T-1}\frac{\eta_t}{B_t}\mathbb{E}[\|\widetilde{g}_t\|_2^2] \le \frac{\Delta_{\mathcal{R}}}{\Lambda_T} + \frac{L_{\mathcal{R}}}{2\Lambda_T}\sum_{t=0}^{T-1}\eta_t^2\mathbb{E}[\mathcal{Q}_t] + \frac{1}{\Lambda_T}\sum_{t=0}^{T-1}\eta_t\mathbb{E}\Big[(\varepsilon_{t,\mathrm{tail}} + \|\xi_t\|_2 + \|e_{\sup}(\Theta_t)\|_2)\sqrt{\mathcal{Q}_t}\Big]. \tag{112}$$

Proof. Equation 107 and the truncation bound imply $\|\widetilde{g}_t - \nabla\mathcal{R}(\Theta_t)\|_2 \le \varepsilon_{t,\mathrm{tail}} + \|\xi_t\|_2 + \|e_{\sup}(\Theta_t)\|_2$. Insert $\widetilde{g}_t$ into the population alignment inner product and apply Lemma 4.1. This gives the pathwise lower bound

$$\langle\nabla\mathcal{R}(\Theta_t), \widehat{g}_t\rangle \ge \frac{1}{B_t}\|\widetilde{g}_t\|_2^2 - (\varepsilon_{t,\mathrm{tail}} + \|\xi_t\|_2 + \|e_{\sup}(\Theta_t)\|_2)\sqrt{\mathcal{Q}_t}.$$

Apply smoothness, sum, and take expectation. The lower bound on $\mathcal{R}$ proves the claim. □

This certificate treats polar reconstruction through its exact positive alignment rather than an absolute polar-error penalty. Its left side measures the retained mean, not the complete population gradient. Full stationarity requires control of the discarded, sampling, and finite-support components. The absolute batch-error term also makes this certificate potentially conservative near stationarity. The conditional-bias theorem provides a complementary route when batch fluctuations are centered.

E.14 Finite-Shot Gradient Error and Adaptive Iterates

Fixed-parameter identity. At a deterministic Θ independent of the support draw, define

$$\mu_c^{\mathrm{pop}}(\Theta) = \mathbb{E}_{x\sim\mathcal{P}_c}[\nabla_\Theta\ell(\Theta; x, c)] \in \mathbb{R}^d,$$
$$\Gamma_{\mathrm{within}}^{\mathrm{pop}}(\Theta) = \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}_{x\sim\mathcal{P}_c}[\|\nabla_\Theta\ell(\Theta; x, c) - \mu_c^{\mathrm{pop}}(\Theta)\|_2^2]. \tag{113}$$

Independence gives the exact identity

$$\mathbb{E}_{\mathcal{D}_{\sup}}[\|e_{\sup}(\Theta)\|_2^2] = \frac{\Gamma_{\mathrm{within}}^{\mathrm{pop}}(\Theta)}{CK}. \tag{114}$$

Proof. At a fixed support-independent Θ, use the same class–shot reindexing as in equation 83. The centered vectors $g_{c,k}(\Theta) - \mu_c^{\mathrm{pop}}(\Theta)$ are independent across the CK observations and have zero expectation. Expanding the squared norm of their average eliminates all cross terms:

$$\mathbb{E}_{\mathcal{D}_{\sup}}[\|e_{\sup}(\Theta)\|_2^2] = \frac{1}{C^2K^2}\sum_{c=1}^{C}\sum_{k=1}^{K}\mathbb{E}[\|g_{c,k}(\Theta) - \mu_c^{\mathrm{pop}}(\Theta)\|_2^2] = \frac{\Gamma_{\mathrm{within}}^{\mathrm{pop}}(\Theta)}{CK}. \tag{115}$$

The identity assumes finite second moments. The $1/(CK)$ factor describes statistical averaging, not an optimization rate. The population dispersion can itself depend on the task, model dimension, and class count. □

Proposition E.19 (A support-uniform gradient bound)**.** *Let $\mathcal{A} \subset \mathbb{R}^d$ be a deterministic compact region independent of the support. On $\mathcal{A}$, assume for each class and almost every image under $\mathcal{P}_c$ that the per-example gradient has norm at most H_{grad} and is L_ℓ-Lipschitz, with finite nonnegative scalar constants. For $\varepsilon_{\mathrm{cov}} > 0$, let $\mathcal{N}_{\mathrm{cov}}$ be the size of a deterministic Euclidean $\varepsilon_{\mathrm{cov}}$-cover of $\mathcal{A}$ with centers in $\mathcal{A}$. For $0 < \delta_{\mathrm{conf}} < 1$, define*

$$\varepsilon_{\mathrm{shot}} = \frac{H_{\mathrm{grad}}}{\sqrt{CK}}\left(1 + \sqrt{2\log\frac{\mathcal{N}_{\mathrm{cov}}}{\delta_{\mathrm{conf}}}}\right) + 2L_\ell\varepsilon_{\mathrm{cov}}. \tag{116}$$

With probability at least $1 - \delta_{\mathrm{conf}}$ over the support draw, $\sup_{\Theta\in\mathcal{A}}\|e_{\sup}(\Theta)\|_2 \le \varepsilon_{\mathrm{shot}}$.

Proof. Fix a cover center in $\mathcal{A}$. The squared-error identity and $\|\nabla \ell\|_2 \le H_{\mathrm{grad}}$ give $\mathbb{E}[\|e_{\mathrm{sup}}(\Theta)\|_2] \le H_{\mathrm{grad}}/\sqrt{CK}$. Replacing one support observation changes the norm of this error by at most $2H_{\mathrm{grad}}/(CK)$. Bounded differences for independent, not necessarily identically distributed observations (McDiarmid, 1989) therefore gives, for $u_{\mathrm{dev}} > 0$,

$$\Pr\left\{ \|e_{\mathrm{sup}}(\Theta)\|_2 > \frac{H_{\mathrm{grad}}}{\sqrt{CK}} + u_{\mathrm{dev}} \right\} \le \exp\left(-\frac{CK\, u_{\mathrm{dev}}^2}{2H_{\mathrm{grad}}^2} \right). \tag{117}$$

For $H_{\mathrm{grad}} > 0$, choose $u_{\mathrm{dev}} = H_{\mathrm{grad}}\sqrt{2\log(\mathcal{N}_{\mathrm{cov}}/\delta_{\mathrm{conf}})/(CK)}$. A union bound controls all $\mathcal{N}_{\mathrm{cov}}$ centers. The zero-gradient case is immediate. The empirical and population data gradients are each L_ℓ-Lipschitz on $\mathcal{A}$. Their difference is therefore $2L_\ell$-Lipschitz. Every point is within $\varepsilon_{\mathrm{cov}}$ of a center, which proves equation 116. The gradient assumptions must hold for population examples almost surely, not just for the observed support. □

Why adaptivity matters. After support-dependent training, conditioning on Θ_t generally changes the support distribution. Equation 117 cannot then be applied as if Θ_t had been fixed before sampling. Proposition E.19 avoids this step by controlling every point in a support-independent region. It covers all learning-rate runs whose trajectories stay in that same region; changing the learning rate does not require a new support draw. It does not establish that the trajectories stay there.

For a deterministic closed ball $\mathcal{A} = \{\Theta : \|\Theta - \Theta_0\|_2 \le R_{\mathrm{loc}}\}$ of scalar radius $R_{\mathrm{loc}} > 0$, a Euclidean cover can satisfy

$$\mathcal{N}_{\mathrm{cov}} \le (1 + 2R_{\mathrm{loc}}/\varepsilon_{\mathrm{cov}})^d. \tag{118}$$

One proof takes a maximal separated set and compares disjoint balls of radius $\varepsilon_{\mathrm{cov}}/2$ with a ball of radius $R_{\mathrm{loc}} + \varepsilon_{\mathrm{cov}}/2$. A maximal separated set is a cover. For full-parameter adaptation, this worst-case bound can be very loose or vacuous. The method does not justify replacing d by r or $\sum_W r_W$: its singular subspaces change with the data and the iteration, and the accumulated parameter displacement need not have low rank. An informative low-dimensional complexity bound would require a separately proved restriction on the entire reachable hypothesis class.

Transfer to support-dependent iterates. On this event, Theorem E.10 implies

$$\frac{1}{\Lambda_T}\sum_{t=0}^{T-1} \eta_t \mathbb{E}[\|\nabla \mathcal{R}(\Theta_t)\|_2^2] \le \frac{4\Delta_F}{\alpha_{\mathrm{olm}}\Lambda_T} + \frac{4\beta_{\mathrm{olm}}}{\alpha_{\mathrm{olm}}} + 2\varepsilon_{\mathrm{shot}}^2 + \frac{2L_F\sigma_{\mathrm{olm}}^2}{\alpha_{\mathrm{olm}}\Lambda_T}\sum_{t=0}^{T-1}\eta_t^2, \tag{119}$$

provided its assumptions hold and the trajectories stay in $\mathcal{A}$ almost surely over training randomness for that support.

Proof. On the uniform support event, every $\Theta \in \mathcal{A}$ satisfies

$$\|\nabla \mathcal{R}(\Theta)\|_2^2 \le 2\|\nabla F(\Theta)\|_2^2 + 2\varepsilon_{\mathrm{shot}}^2. \tag{120}$$

Suppose, for the fixed observed support, the empirical theorem applies and the training trajectories remain in $\mathcal{A}$ almost surely. Average this pointwise inequality with weights η_t/Λ_T, take expectation over training randomness, and insert equation 76. This gives equation 119. No concentration claim is made for support-dependent constants chosen after observing a failed uniform event. □

E.15 A Population-Risk Generalization Bound

For normalized CLIP embeddings and fixed temperature $\tau > 0$, every logit lies in $[-1/\tau, 1/\tau]$. Hence

$$0 \le \ell(\Theta; x, c) \le \log\Big(1 + (C-1)e^{2/\tau}\Big) \le \log C + 2/\tau$$

whenever the normalizations are defined. The bounded-loss premise below therefore follows directly from this classifier. Gradient bounds and parameter Lipschitz constants require separate control of the encoder and normalization maps. Gradient stationarity and prediction risk are different criteria. The next result supplies a risk statement under separate assumptions. It does not infer prediction accuracy from a small gradient.

Proposition E.20 (Uniform risk and optimization–estimation decomposition)**.** *Use the support-independent region and cover of Proposition E.19. Suppose the loss takes values in* $[0, M_{\mathrm{loss}}]$ *and is* H_{grad}*-Lipschitz in* Θ *on this region, for a finite scalar* $M_{\mathrm{loss}} > 0$*. Define the nonnegative scalar*

$$\varepsilon_{\mathrm{risk}} = M_{\mathrm{loss}} \sqrt{\frac{\log(2\mathcal{N}_{\mathrm{cov}}/\delta_{\mathrm{conf}})}{2CK}} + 2H_{\mathrm{grad}}\varepsilon_{\mathrm{cov}}. \tag{121}$$

With probability at least $1 - \delta_{\mathrm{conf}}$ *over the support,*

$$\sup_{\Theta \in \mathcal{A}} |F(\Theta) - \mathcal{R}(\Theta)| \le \varepsilon_{\mathrm{risk}}. \tag{122}$$

On this event, every iterate in $\mathcal{A}$ *obeys*

$$\mathcal{R}(\Theta_T) - \inf_{\Theta \in \mathcal{A}} \mathcal{R}(\Theta) \le F(\Theta_T) - \inf_{\Theta \in \mathcal{A}} F(\Theta) + 2\varepsilon_{\mathrm{risk}}. \tag{123}$$

Proof. At a fixed cover center, Hoeffding's inequality for the CK bounded independent losses (Hoeffding, 1963) gives a two-sided deviation probability at most $2\exp(-2CK\, u_{\mathrm{dev}}^2/M_{\mathrm{loss}}^2)$. Union bound over the cover and extend to $\mathcal{A}$ using the Lipschitz property of both risks. This proves equation 122. Add and subtract the empirical risk at Θ_T and at an arbitrarily accurate minimizer of the population risk over $\mathcal{A}$. Bound the two risk discrepancies by $\varepsilon_{\mathrm{risk}}$, then take the infimum. This proves equation 123. □

The two uniform events for gradients and risks are separate. To require both with failure probability at most δ_{conf}, apply each result with $\delta_{\mathrm{conf}}/2$. A nonconvex stationarity bound alone does not bound the empirical objective-gap term in equation 123. The PL specialization supplies such a route only under its additional assumptions. Neither concentration result establishes a strict advantage over another optimizer without comparing the relevant trajectories, errors, and constants.

E.16 Independent Reference-Gradient Certification

Let $\mathcal{D}_{\mathrm{ref}}$ contain $K_{\mathrm{ref}} \ge 1$ independent images per class, independent of both $\mathcal{D}_{\mathrm{sup}}$ and all training and model-selection decisions that construct the trajectories being audited. Its size is $N_{\mathrm{ref}} = CK_{\mathrm{ref}}$. Let F_{ref} be its empirical objective with the same loss and class weights as equation 1. This is an audit dataset, not extra training data.

Proposition E.21 (Simultaneous reference certificate)**.** *Let* $0 < \delta_{\mathrm{conf}} < 1$*. Condition on a complete training trajectory* $\Theta_0, \ldots, \Theta_{T-1}$ *that is independent of* $\mathcal{D}_{\mathrm{ref}}$*. Suppose per-example gradients on these points have norm at most* H_{grad} *almost surely. Define*

$$\varepsilon_{\mathrm{ref},T} = \frac{H_{\mathrm{grad}}}{\sqrt{N_{\mathrm{ref}}}} \left(1 + \sqrt{2\log(T/\delta_{\mathrm{conf}})}\right). \tag{124}$$

With probability at least $1 - \delta_{\mathrm{conf}}$ *over the reference data, simultaneously for all* $0 \le t < T$*,*

$$\|e_{\mathrm{sup}}(\Theta_t)\|_2 \le \|\nabla F(\Theta_t) - \nabla F_{\mathrm{ref}}(\Theta_t)\|_2 + \varepsilon_{\mathrm{ref},T}. \tag{125}$$

Proof. After conditioning, each parameter value is fixed relative to the independent reference draw. Apply equation 117 to the reference sample with CK replaced by N_{ref}. The union bound over the T fixed points gives $\|\nabla F_{\mathrm{ref}}(\Theta_t) - \nabla \mathcal{R}(\Theta_t)\|_2 \le \varepsilon_{\mathrm{ref},T}$ simultaneously. A triangle inequality proves equation 125. □

For a smaller prespecified list of checkpoints, replace T by its cardinality. No independence between the reference estimates at different checkpoints is required. Independence from trajectory construction is required. Final benchmark test labels must not be used for this audit and then reused as an untouched test set. The inequality is conditional on the stated gradient bound; measuring the reference discrepancy alone does not validate an unknown population concentration constant.

E.17 Stability, Displacement, and Epoch Transfer

Norm control is not global continuity of the merge. Consider one 2×2 tensor, two classes, $K = 1$, $B = 2$, and $r = 1$. For $0 < \vartheta < \pi/2$, take

$$G^{(1)} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad G^{(2),\pm} = \begin{bmatrix} \cos^2 \vartheta & \pm \cos \vartheta \sin \vartheta \\ \sin \vartheta \cos \vartheta & \pm \sin^2 \vartheta \end{bmatrix}. \tag{126}$$

Each matrix has one nonzero singular value equal to one, so truncation is exact. In the positive case the concatenated left and right factors are equal, up to the same harmless column sign changes. In the negative case, they differ by the orthogonal reflection $\mathrm{diag}(1, -1)$ on the right factor's ambient axis. Since the concatenations are square and nonsingular for $\vartheta > 0$, their polar factors are unique and orthogonal. Equation 10 gives exactly

$$\widehat{G}^{+} = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \widehat{G}^{-} = \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{127}$$

As $\vartheta \downarrow 0$, both input pairs converge to the same pair of identical matrices, but the two outputs remain distance one apart in Frobenius norm. The input difference has norm $2 \sin \vartheta$. Therefore, the merge has no uniform Lipschitz constant near this rank-deficient concatenation. Both outputs nevertheless satisfy the same energy identity $\|\widehat{G}^{\pm}\|_F^2 = 1/2$ and the alignment lemma.

This example distinguishes bounded update magnitude from algorithmic stability under data perturbations. Standard SGD stability analyses control distances between neighboring training trajectories (Hardt et al., 2016); they cannot be transferred to this spectral operator without additional regularity conditions. Gaps at the truncation boundary and conditioning of concatenated factors are relevant, but neither is guaranteed by class-unique sampling. The example does not claim that observed CLIP trajectories necessarily encounter this limiting configuration.

What a larger learning-rate interval requires. Under Assumption E.1, conditional expected descent is certified when the alignment and moment are positive and

$$0 < \eta_t < \frac{2\langle \nabla F(\Theta_t), \mathbb{E}_t[\widehat{g}_t] \rangle}{L_F \mathbb{E}_t[\|\widehat{g}_t\|_2^2]}. \tag{128}$$

This follows directly from the smoothness inequality. Compare SGD and OLM at the same Θ_t and under the same conditional batch law. Assume both conditional alignments and both update second moments are strictly positive. The OLM descent interval from this quadratic model is larger precisely when

$$\frac{\langle \nabla F(\Theta_t), \mathbb{E}_t[\widehat{g}_t] \rangle}{\langle \nabla F(\Theta_t), \mathbb{E}_t[g_t] \rangle} > \frac{\mathbb{E}_t[\|\widehat{g}_t\|_2^2]}{\mathbb{E}_t[\|g_t\|_2^2]}. \tag{129}$$

This follows by comparing the two positive ratios in equation 128. The energy lemma controls only part of this comparison. In particular, an upper bound on the OLM moment is not automatically smaller than the SGD moment, which can be small because of cancellation. The constants α_{olm} and M_{olm} can also change with B and r. Consequently, neither a universal B-fold admissible learning-rate gain nor a rank-only optimal-step law follows from the moment bound.

A deterministic reachable region requires a budget assumption. The triangle inequality and the energy bound give

$$\|\Theta_T - \Theta_0\|_2 \le \sum_{t<T} \eta_t \sqrt{\mathcal{Q}_t}. \tag{130}$$

Let $\mathcal{A}$ be the ball in equation 118. Suppose the per-example gradients are bounded by H_{grad} throughout this ball. A sufficient deterministic condition for all iterates through T to remain in the ball is

$$H_{\mathrm{grad}} \sum_{t=0}^{T-1} \frac{\eta_t}{\sqrt{B_t}} \le R_{\mathrm{loc}}. \tag{131}$$

To prove it, induct from Θ_0. At an iterate in the ball, equation 14 bounds the next displacement by $\eta_t H_{\mathrm{grad}}/\sqrt{B_t}$. The sum of preceding bounds keeps the next point inside. The convexity of the ball also contains the update segment. This gives an explicit sufficient region for the uniform statistical theorem without adding projection or clipping to OLM. The condition can be conservative and can fail at long horizons. A support-dependent observed radius cannot be substituted into a support-independent concentration theorem without an additional uniform argument.

Population-loss and accuracy drift. For iteration indices $0 \le t_0 < t_1 \le T$, the same triangle argument gives

$$\|\Theta_{t_1} - \Theta_{t_0}\|_2 \le \sum_{t=t_0}^{t_1-1} \eta_t \sqrt{\mathcal{Q}_t}. \tag{132}$$

If the population loss is H_{grad}-Lipschitz between these points, then

$$|\mathcal{R}(\Theta_{t_1}) - \mathcal{R}(\Theta_{t_0})| \le H_{\mathrm{grad}} \sum_{t=t_0}^{t_1-1} \eta_t \sqrt{\mathcal{Q}_t}. \tag{133}$$

This follows by Lipschitz continuity and equation 132. It controls absolute change, not monotonic improvement.

For an accuracy statement, additionally assume every class probability $p_\Theta(c \mid x)$ is L_{prob}-Lipschitz in Θ on the relevant region, uniformly for population inputs, for a scalar $L_{\mathrm{prob}} > 0$. Define the scalar probability margin

$$\mathsf{m}_\Theta(x, c) = p_\Theta(c \mid x) - \max_{c' \ne c} p_\Theta(c' \mid x) \tag{134}$$

and let $\mathrm{Err}(\Theta)$ be the misclassification probability under uniform class sampling followed by $x \sim \mathcal{P}_c$. Assume zero probability of tied top probabilities at Θ_{t_0}. A previously correct example cannot become incorrect if its initial margin exceeds twice the maximum probability change. Thus

$$\mathrm{Err}(\Theta_{t_1}) - \mathrm{Err}(\Theta_{t_0}) \le \Pr\left\{ 0 < \mathsf{m}_{\Theta_{t_0}}(x, c) \le 2L_{\mathrm{prob}} \sum_{t=t_0}^{t_1-1} \eta_t \sqrt{\mathcal{Q}_t} \right\}. \tag{135}$$

The probability is over a fresh class-balanced query, conditional on the trajectory. This makes the missing assumption explicit: a small parameter displacement protects accuracy only when sufficient classification margin is present. The theorem does not establish such a margin distribution for CLIP.

A conditional certificate for epoch transfer. Let $\mathcal{E}_{\mathrm{eval}} \subseteq \{0, \dots, E\}$ be a finite prespecified epoch grid containing zero. Let $\mathsf{Risk}_{\mathrm{src}}(e)$ and $\mathsf{Risk}_{\mathrm{tgt}}(e)$ be the scalar population evaluation risks of separately trained source- and target-task trajectories at epoch e. They may be misclassification risks or the same chosen bounded loss. Let $\widehat{\mathsf{Risk}}_{\mathrm{src}}(e)$ be an independent-validation estimate. Suppose, for nonnegative scalars $\varepsilon_{\mathrm{val}}$ and Δ_{curve},

$$\sup_{e \in \mathcal{E}_{\mathrm{eval}}} |\widehat{\mathsf{Risk}}_{\mathrm{src}}(e) - \mathsf{Risk}_{\mathrm{src}}(e)| \le \varepsilon_{\mathrm{val}}, \tag{136}$$

$$\sup_{e \in \mathcal{E}_{\mathrm{eval}}} |[\mathsf{Risk}_{\mathrm{tgt}}(e) - \mathsf{Risk}_{\mathrm{tgt}}(0)] - [\mathsf{Risk}_{\mathrm{src}}(e) - \mathsf{Risk}_{\mathrm{src}}(0)]| \le \Delta_{\mathrm{curve}}. \tag{137}$$

For any source-selected epoch $\widehat{e} \in \arg\min_{e \in \mathcal{E}_{\mathrm{eval}}} \widehat{\mathsf{Risk}}_{\mathrm{src}}(e)$,

$$\mathsf{Risk}_{\mathrm{tgt}}(\widehat{e}) - \min_{e \in \mathcal{E}_{\mathrm{eval}}} \mathsf{Risk}_{\mathrm{tgt}}(e) \le 2\varepsilon_{\mathrm{val}} + 2\Delta_{\mathrm{curve}}. \tag{138}$$

To prove the statement, compare the target risk at $\widehat{e}$ with its grid minimizer. Equation 137 replaces their difference by the corresponding source-risk difference with error at most $2\Delta_{\mathrm{curve}}$. Source validation optimality and equation 136 bound that difference by $2\varepsilon_{\mathrm{val}}$.

This is a conditional transfer statement, not evidence that OLM makes two tasks' risk profiles similar. Centering removes irrelevant task-specific baseline offsets. The comparison is over the specified grid, not every possible epoch. A broad target plateau can also make an epoch choice inexpensive, but it must be demonstrated on the target tasks. No assertion of universal epoch transfer follows from within-task stationarity or bounded update energy.

E.18 Numerical Reconstruction Errors

Proposition E.22 (Perturbing the polar factors)**.** *For fixed retained factors and coefficients, suppose the numerical polar factors U_{num}, V_{num} satisfy $\|U_{\mathrm{num}} - U_{\mathrm{pol}}\|_2 \le a_U$ and $\|V_{\mathrm{num}} - V_{\mathrm{pol}}\|_2 \le a_V$.*

Set $z_W = (a_U + a_V + a_U a_V)\|s\|_2$ *for this tensor. Then the numerical reconstruction* $G_{\mathrm{num}} = U_{\mathrm{num}} \operatorname{diag}(s) V_{\mathrm{num}}^\top$ *satisfies*

$$\|G_{\mathrm{num}} - \widehat{G}\|_F \le z_W.$$

After concatenating tensors, let $z_t = (\sum_W z_W^2)^{1/2}$. *Then*

$$\begin{aligned} \|g_{t,\mathrm{num}}\|_2 &\le \sqrt{\mathcal{Q}_t} + z_t, \\ \langle \widetilde{g}_t, g_{t,\mathrm{num}} \rangle &\ge \sum_W \frac{\|\widetilde{G}_t^W\|_F^2}{\|U_t^W\|_2 \|V_t^W\|_2} - \|\widetilde{g}_t\|_2 z_t. \end{aligned} \tag{139}$$

Proof. Expand the reconstruction difference into its left perturbation, right perturbation, and product perturbation. Bound their Frobenius norms by $a_U\|s\|_2$, $a_V\|s\|_2$, and $a_U a_V\|s\|_2$, respectively. Sum across tensors in squared norm. The triangle and Cauchy–Schwarz inequalities and the exact geometric lemma give equation 139. □

This perturbation statement compares numerical polar factors to the selected exact factors of the same retained inputs. It is not a Lipschitz bound for the polar map as its input changes. Rank-deficient inputs need not admit such a uniform bound. Approximate gradient SVDs must separately report their actual retained reconstructions and truncation residuals. Casting the reconstructed update or applying the parameter update at lower precision introduces an additional measured update error, which can be added to z_t.

E.19 A Smooth Strongly Convex Counterexample

The following example, within the general smooth-objective setting of the theorem, shows why Assumption E.2 cannot be removed. Use one 2×2 tensor W and let $\Theta = \mathrm{vec}(W) \in \mathbb{R}^4$. Take $C = 3$, $K = 1$, and a class-unique batch containing all three examples. Let $r = r_W = 1$ and choose a scalar $\kappa \in (0, 1)$. Suppressing iteration and tensor indices, the gradient matrices at $W = 0$ are

$$G^{(1)} = \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix}, \qquad G^{(2)} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \qquad G^{(3)} = \begin{bmatrix} -3 & 0 \\ 0 & -(2+\kappa) \end{bmatrix}. \tag{140}$$

The matrix representation of the ordinary mean gradient is

$$\overline{G} = \frac{1}{3}\sum_{b=1}^{3} G^{(b)} = \begin{bmatrix} 0 & 0 \\ 0 & -\kappa/3 \end{bmatrix}. \tag{141}$$

The top singular direction of $G^{(3)}$ is its first coordinate because $3 > 2 + \kappa$. Direct substitution into equations 6–10 gives

$$\widehat{G} = \begin{bmatrix} 0 & 0 \\ 0 & 2/3 \end{bmatrix}. \tag{142}$$

Therefore,

$$\langle \overline{G}, \widehat{G} \rangle_F = -\frac{2\kappa}{9} < 0. \tag{143}$$

To realize these matrices as gradients, define

$$\ell_b(\Theta) = \frac{1}{2}\|W + G^{(b)}\|_F^2, \qquad F(\Theta) = \frac{1}{3}\sum_{b=1}^{3} \ell_b(\Theta). \tag{144}$$

At $\Theta = 0$, the per-sample gradients are exactly the matrices in equation 140. The function F is 1-smooth and 1-strongly convex. Nevertheless, for every $\eta > 0$,

$$F(-\eta\,\mathrm{vec}(\widehat{G})) - F(0) = -\eta\langle \overline{G}, \widehat{G} \rangle_F + \frac{\eta^2}{2}\|\widehat{G}\|_F^2 > 0. \tag{145}$$

Thus, an OLM step is an ascent step at this point. When $\kappa = 0$, the true mean gradient is zero but $\widehat{G} \neq 0$. This also shows that rank truncation can create a nonvanishing additive bias at a stationary point.

E.20 No Distribution-Free Exact Population Guarantee at Fixed Shot

Proposition E.23 (Two-class cross-entropy obstruction)**.** *For every finite $K \geq 1$, a method that observes only a balanced support set with K examples per class and its internal randomness cannot guarantee population stationarity almost surely for every class-conditional distribution. The obstruction already holds for $C = 2$, one trainable scalar, and a smooth convex cross-entropy objective. It also precludes a distribution-free almost-sure vanishing weighted-average population-gradient norm along the trajectory.*

Proof. Take three images of the existing input dimensions and a fixed scalar feature ϕ that takes values $0, +1, -1$ on them. Use one scalar parameter $\Theta \in \mathbb{R}$ and the two-class probabilities

$$p_\Theta(1 \mid x) = \frac{\exp(\Theta\phi(x))}{1 + \exp(\Theta\phi(x))}, \qquad p_\Theta(2 \mid x) = \frac{1}{1 + \exp(\Theta\phi(x))}. \tag{146}$$

Under the positive problem, both classes have feature zero with probability $1/2$. Conditional on a nonzero feature, class 1 takes $+1$ with probability $3/4$ and -1 otherwise; class 2 reverses these signs. Under the negative problem, reverse the nonzero signs again. Give both problems the same initialization and internal random seed. Their population risks are

$$\mathcal{R}^{\pm}(\Theta) = \frac{1}{2}\log 2 + \frac{1}{2}\left[p_{\pm}\log(1 + e^{-\Theta}) + (1 - p_{\pm})\log(1 + e^{\Theta})\right], \quad p_+ = \tfrac{3}{4},\ p_- = \tfrac{1}{4}. \tag{147}$$

They are strictly convex and $1/8$-smooth, with unique stationary points $+\log 3$ and $-\log 3$. Their derivatives differ by a constant:

$$\frac{d}{d\Theta}\mathcal{R}^{+}(\Theta) - \frac{d}{d\Theta}\mathcal{R}^{-}(\Theta) = -\frac{1}{4}. \tag{148}$$

Thus, at every common parameter value,

$$\left|\frac{d}{d\Theta}\mathcal{R}^{+}(\Theta)\right|^2 + \left|\frac{d}{d\Theta}\mathcal{R}^{-}(\Theta)\right|^2 \geq \frac{1}{32}. \tag{149}$$

The event that all $2K$ support features equal zero has probability $2^{-2K} > 0$ under both problems. On that event, the labeled supports are identical. Conditional on a common internal random seed, every support-only algorithm produces the same trajectory in the two problems. Equation 149 rules out a stationary output in both. Applying the same inequality at every iterate and averaging with the deterministic weights η_t/Λ_T rules out vanishing weighted-average squared gradients in both problems. The absolute-gradient version follows from the derivative separation. This is a general cross-entropy example, not a claim that the specific CLIP architecture equals this scalar classifier. □

The proposition is an information limitation, not an OLM-specific failure. It justifies retaining a finite-shot discrepancy or imposing a separate population-alignment assumption. It is not a tight minimax sample-complexity rate. Structural assumptions on the task or additional independent data can yield sharper positive results.

F Comparison against State-of-the-Art PEFT Methods

This appendix examines four complementary aspects of few-shot adaptation: test accuracy, transfer of the training duration, long-horizon dynamics, and degradation after the peak performance on test data. We compare OLM-SGD with 13 few-shot adaptation baselines in Tables 20–22. These include parameter-efficient methods and frozen-feature alternatives. The trajectory analyses focus on LoRA (Hu et al., 2022), SO (Mrabah et al., 2025), and SAFT (Nguyen et al., 2025). LoRA uses rank 2. SO and SAFT use 99.9% sparsity. OLM-SGD uses gradient rank $r = 3$.

F.1 Few-Shot Accuracy across Backbones and Shot Levels

Protocol. Tables 20, 21, and 22 report results on all 11 datasets with ViT-B/32, ViT-B/16, and ViT-L/14. We evaluate $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class using seeds 1, 2, and 3. Entries report mean top-1 accuracy and population standard deviation across seeds. The Average column first averages the 11 datasets within each seed. It then reports the mean and standard deviation of these three averages. The method-specific schedules follow Appendix D.

Table 19: Summary of the few-shot comparison. Each entry gives OLM-SGD's mean accuracy (%) and, in parentheses, its margin over the highest baseline mean at the same backbone and shot count. Both use the equally weighted 11-dataset averages reported in Tables 20–22. Margins are computed from the displayed one-decimal means. The full tables report seed variability.

Backbone	1 shot	2 shots	4 shots	8 shots	16 shots
ViT-B/32	67.2 (+0.9)	70.4 (+1.7)	73.1 (+1.8)	75.8 (+1.7)	78.5 (+1.5)
ViT-B/16	72.0 (+1.2)	75.1 (+1.6)	77.8 (+1.8)	80.3 (+1.9)	82.6 (+1.5)
ViT-L/14	77.4 (+0.1)	80.1 (+0.1)	82.5 (+0.4)	84.7 (+0.6)	86.8 (+0.8)

Average accuracy across the 15 settings. OLM-SGD attains the highest reported average accuracy in every backbone–shot setting. Table 19 summarizes these results. On ViT-B/32, its mean accuracy increases from 67.2% at 1 shot to 78.5% at 16 shots. The margin over the strongest baseline at each shot count ranges from 0.9 to 1.8 percentage points. On ViT-B/16, accuracy increases from 72.0% to 82.6%, with margins of 1.2–1.9 points. The advantage is smaller on ViT-L/14. Its accuracy increases from 77.4% to 86.8%, with margins of 0.1–0.8 points. Averaging the five reported margins gives 1.52, 1.60, and 0.40 points for the three backbones, respectively. Thus, the improvement persists across model scales but is strongest for the ViT-B backbones.

Dataset-level strengths and exceptions. Using the displayed means, OLM-SGD is best or tied for best in $32/55$, $37/55$, and $25/55$ dataset–shot settings for ViT-B/32, ViT-B/16, and ViT-L/14, respectively. Across all three backbones, it is best or tied in $14/15$ settings on SUN397, $13/15$ on Aircraft, and $14/15$ on Stanford Cars. For example, 16-shot ViT-B/16 accuracy on Aircraft is 52.9%, compared with 48.3% for the strongest baseline, GDA. At 4 shots on Stanford Cars, OLM-SGD reaches 76.5%, compared with 72.7% for PiSSA. These results show that the mean improvement is supported by repeated gains on several datasets.

The advantage is not uniform. OLM-SGD remains below the strongest ImageNet baseline at every shot count for both ViT-B backbones. EuroSAT also provides counterexamples. At 2 shots with ViT-L/14, SAFT reaches 85.1%, whereas OLM-SGD obtains 79.6%. These exceptions distinguish the strongest overall average from dominance on every task.

Table 20: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	ProLIP	63.4 ± 0.1	65.7 ± 0.3	22.7 ± 0.5	60.3 ± 4.3	59.6 ± 0.9	79.5 ± 0.2	86.6 ± 0.4	74.4 ± 0.9	93.2 ± 0.5	50.8 ± 0.3	69.1 ± 0.5	65.9 ± 0.3
	Adapter	59.7 ± 0.1	63.9 ± 0.2	22.1 ± 0.5	56.8 ± 2.9	58.6 ± 1.2	78.6 ± 0.5	85.3 ± 0.7	72.4 ± 0.9	92.9 ± 0.6	49.1 ± 0.6	68.6 ± 0.4	64.4 ± 0.2
	SAFT	59.7 ± 0.7	64.4 ± 0.2	21.0 ± 1.0	66.2 ± 1.8	57.5 ± 1.0	78.0 ± 0.6	85.3 ± 1.1	73.3 ± 0.7	93.3 ± 0.7	47.6 ± 0.9	69.7 ± 0.6	65.1 ± 0.1
	CoOp	58.6 ± 0.3	61.0 ± 0.4	20.6 ± 0.4	55.9 ± 5.8	58.5 ± 0.4	77.0 ± 1.0	85.6 ± 0.7	72.1 ± 1.4	89.9 ± 1.4	44.3 ± 0.5	64.9 ± 2.0	62.6 ± 0.3
	KgCoOp	64.3 ± 0.1	66.1 ± 0.1	21.4 ± 0.1	55.4 ± 4.8	60.8 ± 0.2	80.7 ± 0.1	87.9 ± 0.2	67.4 ± 0.6	92.8 ± 0.1	48.1 ± 1.2	67.5 ± 0.4	64.8 ± 0.4
	ProGrad	59.0 ± 0.9	62.9 ± 1.0	15.6 ± 4.8	54.9 ± 4.7	59.0 ± 0.6	78.2 ± 1.1	83.8 ± 2.4	69.0 ± 0.4	90.0 ± 1.4	43.0 ± 1.3	64.7 ± 0.8	61.8 ± 0.5
	TIP-Adapter	62.2 ± 0.0	62.4 ± 0.0	18.7 ± 0.2	38.0 ± 0.9	60.8 ± 0.1	78.9 ± 0.0	85.1 ± 0.1	65.4 ± 0.2	91.5 ± 0.0	42.9 ± 0.0	61.9 ± 0.0	60.7 ± 0.1
	GDA	62.4 ± 0.0	36.7 ± 0.8	15.8 ± 1.3	48.9 ± 4.4	27.5 ± 1.5	35.5 ± 1.2	34.2 ± 0.7	59.4 ± 1.5	74.2 ± 2.7	29.5 ± 1.7	47.3 ± 1.8	42.8 ± 0.2
	ProKeR	62.7 ± 0.1	64.5 ± 0.1	21.5 ± 0.4	52.5 ± 5.3	62.4 ± 0.5	79.1 ± 0.1	85.3 ± 0.3	72.4 ± 0.8	92.1 ± 0.3	47.2 ± 0.5	65.3 ± 0.4	64.1 ± 0.5
	LoRA	57.4 ± 0.0	64.6 ± 0.4	22.2 ± 0.6	62.3 ± 4.4	60.1 ± 0.5	78.9 ± 0.3	89.0 ± 0.7	72.9 ± 0.8	93.7 ± 0.7	49.8 ± 0.2	69.3 ± 0.6	65.5 ± 0.3
	DoRA	57.3 ± 0.1	65.0 ± 0.5	21.9 ± 0.3	63.3 ± 5.8	59.9 ± 0.4	79.3 ± 0.4	89.4 ± 0.7	74.8 ± 1.0	93.6 ± 0.6	50.0 ± 0.1	69.9 ± 0.3	65.9 ± 0.5
	PiSSA	56.3 ± 0.6	66.0 ± 0.3	20.4 ± 0.9	61.2 ± 4.7	58.3 ± 0.4	77.7 ± 0.4	85.8 ± 1.6	75.9 ± 1.4	93.1 ± 0.6	47.6 ± 1.6	70.6 ± 0.3	64.8 ± 0.4
	SO	60.0 ± 0.2	65.6 ± 0.2	22.4 ± 0.1	60.4 ± 4.9	60.8 ± 0.4	80.5 ± 0.2	89.0 ± 0.6	76.7 ± 0.5	93.6 ± 0.4	50.1 ± 0.5	70.3 ± 0.3	66.3 ± 0.3
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.6	50.7 ± 1.9	70.1 ± 0.4	67.2 ± 0.5
2	ProLIP	63.6 ± 0.1	67.1 ± 0.3	23.6 ± 0.7	71.6 ± 0.5	63.2 ± 0.3	79.8 ± 0.1	87.2 ± 0.8	78.5 ± 1.0	93.4 ± 0.5	54.4 ± 0.9	72.3 ± 0.4	68.6 ± 0.1
	Adapter	60.6 ± 0.2	65.7 ± 0.2	23.1 ± 0.2	67.7 ± 0.7	61.5 ± 0.5	79.1 ± 0.6	87.0 ± 0.3	77.7 ± 1.1	93.9 ± 0.3	54.0 ± 0.4	71.2 ± 0.4	67.4 ± 0.1
	SAFT	61.3 ± 0.1	65.2 ± 0.3	23.2 ± 0.6	78.1 ± 2.5	58.8 ± 0.0	78.8 ± 0.2	89.2 ± 0.2	78.6 ± 1.1	93.9 ± 0.3	53.8 ± 1.4	71.3 ± 0.2	68.4 ± 0.4
	CoOp	61.0 ± 0.4	63.7 ± 0.3	21.5 ± 0.6	67.9 ± 1.8	59.6 ± 0.5	76.9 ± 1.1	87.1 ± 1.0	76.5 ± 0.3	93.0 ± 0.2	47.2 ± 1.5	70.2 ± 0.6	65.9 ± 0.4
	KgCoOp	64.5 ± 0.1	67.0 ± 0.2	22.2 ± 0.2	60.1 ± 1.3	61.3 ± 0.3	80.9 ± 0.1	88.9 ± 0.5	68.7 ± 1.3	93.0 ± 0.2	49.9 ± 0.9	68.9 ± 0.5	65.9 ± 0.4
	ProGrad	59.0 ± 1.2	65.0 ± 1.0	18.5 ± 4.4	62.8 ± 2.9	56.9 ± 7.5	78.6 ± 0.6	87.4 ± 0.3	73.5 ± 2.2	92.2 ± 0.3	46.7 ± 2.5	68.3 ± 1.0	64.5 ± 0.8
	TIP-Adapter	62.2 ± 0.1	62.8 ± 0.0	20.2 ± 0.1	40.5 ± 0.5	61.7 ± 0.1	79.0 ± 0.0	85.3 ± 0.0	70.4 ± 0.6	91.6 ± 0.2	43.1 ± 0.1	62.2 ± 0.1	61.7 ± 0.1
	GDA	62.9 ± 0.0	64.5 ± 0.2	22.0 ± 0.6	66.7 ± 1.0	63.0 ± 0.3	79.0 ± 0.2	85.0 ± 0.0	83.2 ± 2.0	92.9 ± 0.7	48.5 ± 3.1	68.0 ± 0.6	66.9 ± 0.5
	ProKeR	63.4 ± 0.1	66.1 ± 0.1	23.0 ± 0.2	60.5 ± 1.5	64.3 ± 0.3	79.4 ± 0.1	85.4 ± 0.4	75.5 ± 0.6	93.3 ± 0.3	50.3 ± 0.4	67.5 ± 0.3	66.2 ± 0.1
	LoRA	58.3 ± 0.2	65.8 ± 0.7	24.3 ± 0.3	68.5 ± 1.6	60.1 ± 0.7	79.3 ± 0.3	89.4 ± 0.6	78.6 ± 1.1	94.3 ± 0.2	53.1 ± 1.7	71.8 ± 0.4	67.6 ± 0.2
	DoRA	58.2 ± 0.3	66.0 ± 0.5	24.4 ± 0.2	72.3 ± 2.5	59.9 ± 0.7	79.7 ± 0.2	89.6 ± 0.6	79.5 ± 0.8	94.3 ± 0.1	53.6 ± 1.7	72.2 ± 0.3	68.2 ± 0.1
	PiSSA	57.8 ± 0.5	67.2 ± 0.4	22.1 ± 0.2	74.3 ± 2.9	60.6 ± 0.6	78.8 ± 0.2	88.2 ± 0.5	81.4 ± 0.4	93.9 ± 0.2	53.7 ± 1.1	73.1 ± 0.4	68.3 ± 0.3
	SO	58.7 ± 0.3	66.7 ± 0.3	24.0 ± 0.5	71.6 ± 1.4	60.2 ± 0.1	80.1 ± 0.2	89.5 ± 0.6	82.3 ± 1.0	94.2 ± 0.2	55.5 ± 0.9	73.0 ± 0.3	68.7 ± 0.0
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.9 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.4 ± 0.9	70.4 ± 0.1
4	ProLIP	64.6 ± 0.2	68.7 ± 0.5	27.1 ± 0.5	75.8 ± 1.7	65.1 ± 0.4	80.2 ± 0.3	87.9 ± 0.8	84.7 ± 0.4	94.5 ± 0.5	58.6 ± 0.5	74.4 ± 0.8	71.1 ± 0.3
	Adapter	61.5 ± 0.3	67.2 ± 0.5	25.1 ± 0.3	70.8 ± 1.5	63.9 ± 0.4	79.1 ± 0.2	87.2 ± 0.9	84.6 ± 0.5	94.4 ± 0.2	59.2 ± 0.8	74.4 ± 0.5	69.8 ± 0.3
	SAFT	62.6 ± 0.4	66.0 ± 0.3	24.6 ± 0.5	83.1 ± 1.6	61.8 ± 0.4	79.4 ± 0.4	90.4 ± 0.1	83.4 ± 1.0	94.7 ± 0.2	56.7 ± 0.3	73.5 ± 0.3	70.6 ± 0.1
	CoOp	63.2 ± 0.2	66.0 ± 0.1	23.7 ± 1.2	74.0 ± 1.4	62.9 ± 0.5	76.9 ± 0.3	87.8 ± 1.0	83.4 ± 1.9	93.0 ± 0.6	54.4 ± 1.4	71.6 ± 0.5	68.8 ± 0.2
	KgCoOp	64.9 ± 0.0	67.8 ± 0.1	22.7 ± 0.1	60.7 ± 2.6	61.9 ± 0.3	81.1 ± 0.2	89.0 ± 0.3	70.3 ± 0.2	93.6 ± 0.1	52.9 ± 0.5	69.6 ± 0.3	66.8 ± 0.4
	ProGrad	62.8 ± 0.5	67.7 ± 0.5	24.2 ± 0.8	68.5 ± 0.8	63.4 ± 0.9	79.1 ± 0.2	87.6 ± 0.6	79.2 ± 1.2	75.9 ± 23.6	51.9 ± 1.2	70.8 ± 0.3	66.4 ± 2.0
	TIP-Adapter	61.7 ± 0.2	63.5 ± 0.2	23.3 ± 0.4	45.2 ± 0.9	63.1 ± 0.1	79.2 ± 0.0	85.9 ± 0.1	82.6 ± 0.3	92.3 ± 0.2	43.7 ± 0.3	63.3 ± 0.2	64.0 ± 0.2
	GDA	64.7 ± 0.1	67.7 ± 1.6	26.2 ± 1.1	74.1 ± 2.0	65.1 ± 0.3	79.5 ± 0.2	84.7 ± 1.0	90.5 ± 0.5	94.2 ± 0.2	55.7 ± 1.3	72.8 ± 0.6	70.5 ± 0.2
	ProKeR	64.7 ± 0.1	67.9 ± 0.1	25.3 ± 0.3	66.3 ± 2.8	66.5 ± 0.2	80.1 ± 0.1	85.9 ± 0.1	78.3 ± 0.2	94.4 ± 0.1	54.0 ± 0.2	69.9 ± 0.4	68.5 ± 0.2
	LoRA	60.0 ± 0.1	67.3 ± 0.3	26.7 ± 0.7	71.9 ± 1.8	63.0 ± 0.6	79.4 ± 0.4	90.3 ± 0.2	84.2 ± 1.4	95.1 ± 0.1	57.0 ± 0.7	74.4 ± 0.7	69.9 ± 0.2
	DoRA	59.8 ± 0.0	67.7 ± 0.3	26.2 ± 0.4	77.1 ± 2.2	63.4 ± 0.7	79.8 ± 0.3	90.2 ± 0.2	84.5 ± 1.4	94.9 ± 0.2	57.5 ± 0.8	74.6 ± 0.5	70.5 ± 0.2
	PiSSA	58.1 ± 0.5	68.4 ± 0.2	26.2 ± 0.4	80.5 ± 0.9	64.9 ± 0.7	80.0 ± 0.2	90.3 ± 0.6	85.9 ± 0.9	94.8 ± 0.2	59.6 ± 0.3	75.6 ± 0.5	71.3 ± 0.1
	SO	57.2 ± 0.0	68.0 ± 0.2	25.8 ± 0.4	76.0 ± 2.7	62.9 ± 0.3	80.2 ± 0.2	90.0 ± 0.6	88.0 ± 0.4	94.7 ± 0.2	59.7 ± 0.3	75.0 ± 0.6	70.7 ± 0.3
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.1 ± 0.6	81.4 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1 ± 0.2
8	ProLIP	65.1 ± 0.1	70.1 ± 0.2	31.1 ± 0.7	82.2 ± 0.7	67.9 ± 0.2	79.3 ± 1.3	86.3 ± 3.6	87.6 ± 0.2	94.7 ± 0.1	62.8 ± 1.0	76.2 ± 0.9	73.0 ± 0.5
	Adapter	62.5 ± 0.1	68.8 ± 0.1	29.3 ± 0.6	77.3 ± 0.9	68.2 ± 0.6	79.1 ± 0.1	88.3 ± 0.7	88.7 ± 0.2	94.5 ± 0.1	63.1 ± 0.7	77.2 ± 0.2	72.5 ± 0.1
	SAFT	64.2 ± 0.2	67.7 ± 0.3	28.9 ± 0.2	88.0 ± 2.0	65.3 ± 0.4	79.4 ± 0.4	90.5 ± 0.3	87.4 ± 0.5	94.8 ± 0.5	60.9 ± 1.1	75.7 ± 0.5	73.0 ± 0.2
	CoOp	65.0 ± 0.1	68.6 ± 0.2	27.9 ± 0.6	77.5 ± 0.8	66.8 ± 0.8	77.0 ± 0.6	87.9 ± 1.1	88.3 ± 0.6	93.1 ± 0.7	59.7 ± 0.5	75.5 ± 0.4	71.6 ± 0.0
	KgCoOp	65.2 ± 0.0	68.6 ± 0.2	23.8 ± 0.2	67.4 ± 1.7	62.6 ± 0.1	81.5 ± 0.1	89.9 ± 0.3	73.3 ± 1.0	93.8 ± 0.1	56.3 ± 1.3	71.8 ± 0.6	68.6 ± 0.4
	ProGrad	57.4 ± 8.8	69.4 ± 0.2	25.4 ± 1.5	72.3 ± 2.6	66.7 ± 0.3	79.0 ± 0.7	88.8 ± 0.2	81.2 ± 1.2	93.0 ± 0.3	56.7 ± 1.2	71.4 ± 0.4	69.2 ± 0.7
	TIP-Adapter	61.7 ± 0.2	65.0 ± 0.0	27.1 ± 1.0	63.2 ± 1.1	65.8 ± 0.3	79.8 ± 0.0	87.1 ± 0.3	89.0 ± 0.3	92.6 ± 0.1	46.5 ± 0.3	65.3 ± 0.2	67.6 ± 0.1
	GDA	65.8 ± 0.1	70.6 ± 0.2	33.2 ± 0.7	81.5 ± 2.0	70.8 ± 0.3	79.3 ± 0.2	85.2 ± 0.2	92.9 ± 0.3	94.0 ± 0.1	61.9 ± 1.3	77.4 ± 0.5	73.9 ± 0.2
	ProKeR	65.1 ± 0.1	70.5 ± 0.2	28.3 ± 0.4	77.5 ± 1.2	70.7 ± 0.2	80.7 ± 0.1	86.8 ± 0.6	82.8 ± 0.2	94.7 ± 0.1	59.1 ± 0.5	73.4 ± 0.2	71.8 ± 0.2
	LoRA	62.1 ± 0.1	69.4 ± 0.3	31.4 ± 0.6	85.2 ± 1.7	68.2 ± 0.6	79.3 ± 0.3	90.3 ± 0.2	88.8 ± 0.9	94.7 ± 0.1	62.3 ± 0.0	76.7 ± 0.2	73.5 ± 0.1
	DoRA	62.0 ± 0.1	69.8 ± 0.4	31.2 ± 0.4	86.9 ± 1.0	68.3 ± 0.6	79.6 ± 0.2	90.6 ± 0.2	88.9 ± 0.7	94.7 ± 0.3	62.4 ± 0.5	77.0 ± 0.4	73.8 ± 0.1
	PiSSA	60.1 ± 0.3	70.2 ± 0.1	30.8 ± 0.5	88.9 ± 0.4	68.8 ± 0.4	80.3 ± 0.2	90.7 ± 0.3	89.5 ± 0.5	94.9 ± 0.4	63.1 ± 0.3	77.8 ± 0.1	74.1 ± 0.1
	SO	55.5 ± 0.2	69.6 ± 0.1	30.9 ± 0.9	87.4 ± 0.3	67.7 ± 0.3	80.1 ± 0.1	90.4 ± 0.2	91.2 ± 0.3	94.8 ± 0.3	65.4 ± 0.5	77.3 ± 0.3	73.7 ± 0.1
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.6 ± 0.1	91.2 ± 0.5	91.9 ± 0.1	95.2 ± 0.2	64.6 ± 0.6	78.5 ± 0.4	75.8 ± 0.1
16	ProLIP	66.2 ± 0.2	71.2 ± 0.1	35.2 ± 1.0	85.7 ± 0.4	71.4 ± 0.8	80.1 ± 0.5	89.2 ± 0.7	89.1 ± 3.7	94.7 ± 0.3	67.2 ± 1.1	78.7 ± 0.6	75.3 ± 0.7
	Adapter	64.4 ± 0.1	70.7 ± 0.1	33.7 ± 0.3	84.2 ± 1.1	72.0 ± 0.4	79.5 ± 0.1	88.4 ± 0.4	93.1 ± 0.1	95.3 ± 0.1	67.3 ± 0.9	79.2 ± 0.4	75.2 ± 0.2
	SAFT	66.3 ± 0.2	69.2 ± 0.5	35.4 ± 0.3	92.3 ± 0.7	69.9 ± 0.2	79.7 ± 0.3	90.6 ± 0.1	91.1 ± 0.3	95.3 ± 0.1	65.2 ± 0.5	78.2 ± 0.4	75.8 ± 0.1
	CoOp	66.6 ± 0.3	71.0 ± 0.4	31.1 ± 0.2	83.2 ± 1.2	70.9 ± 0.5	78.4 ± 0.4	88.0 ± 0.3	92.4 ± 0.2	94.2 ± 0.3	63.5 ± 0.4	77.7 ± 0.5	74.3 ± 0.1
	KgCoOp	65.5 ± 0.1	69.9 ± 0.1	25.6 ± 0.4	70.0 ± 1.2	63.9 ± 0.1	81.6 ± 0.1	90.4 ± 0.1	79.4 ± 0.9	94.0 ± 0.1	59.9 ± 1.0	73.6 ± 0.1	70.4 ± 0.2
	ProGrad	53.8 ± 6.5	71.3 ± 0.1	29.9 ± 1.0	77.3 ± 0.7	70.2 ± 0.6	80.1 ± 0.4	89.2 ± 0.9	86.4 ± 0.6	93.6 ± 0.2	60.8 ± 1.3	74.7 ± 0.3	71.6 ± 0.5
	TIP-Adapter	62.7 ± 0.4	68.1 ± 0.1	32.4 ± 0.4	80.0 ± 0.3	70.4 ± 0.4	80.5 ± 0.1	88.1 ± 0.8	92.9 ± 0.2	93.5 ± 0.2	51.8 ± 0.5	68.5 ± 0.2	71.7 ± 0.0
	GDA	66.3 ± 0.0	73.5 ± 0.2	40.2 ± 0.4	86.5 ± 0.7	76.0 ± 0.3	80.9 ± 0.1	87.2 ± 1.6	95.8 ± 0.1	94.8 ± 0.2	66.3 ± 0.1	79.8 ± 0.2	77.0 ± 0.1
	ProKeR	66.9 ± 0.0	73.3 ± 0.1	33.5 ± 0.3	84.0 ± 0.2	75.6 ± 0.2	81.4 ± 0.1	87.6 ± 0.3	87.5 ± 0.1	95.6 ± 0.1	65.3 ± 0.7	77.1 ± 0.3	75.3 ± 0.1
	LoRA	64.9 ± 0.3	71.3 ± 0.2	38.4 ± 1.2	91.2 ± 0.8	73.0 ± 0.2	79.6 ± 0.2	90.5 ± 0.2	92.9 ± 0.2	95.2 ± 0.2	66.8 ± 0.3	79.0 ± 0.3	76.6 ± 0.2
	DoRA	64.8 ± 0.2	71.5 ± 0.2	38.8 ± 0.6	91.3 ± 0.7	73.5 ± 0.2	79.9 ± 0.1	90.6 ± 0.2	92.9 ± 0.1	95.2 ± 0.3	67.0 ± 0.3	79.4 ± 0.2	76.8 ± 0.1
	PiSSA	63.4 ± 0.2	71.6 ± 0.2	37.1 ± 0.6	92.0 ± 0.6	73.1 ± 0.3	80.5 ± 0.1	90.8 ± 0.2	93.3 ± 0.0	95.6 ± 0.3	67.8 ± 0.3	79.4 ± 0.2	76.8 ± 0.1
	SO	52.7 ± 0.2	70.8 ± 0.1	36.8 ± 0.6	91.9 ± 0.6	72.0 ± 0.3	79.9 ± 0.0	90.7 ± 0.2	94.6 ± 0.1	95.6 ± 0.1	69.0 ± 0.6	79.6 ± 0.2	75.8 ± 0.1
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5 ± 0.1

Table 21: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	ProLIP	68.7 ± 0.3	68.2 ± 0.2	29.3 ± 0.7	65.0 ± 4.1	66.8 ± 0.1	85.4 ± 0.4	90.7 ± 0.3	78.3 ± 0.2	94.2 ± 0.2	52.5 ± 1.4	74.1 ± 0.6	70.3 ± 0.4
	Adapter	65.8 ± 0.1	67.0 ± 0.2	28.1 ± 0.5	62.5 ± 4.7	66.8 ± 0.2	84.9 ± 0.2	90.0 ± 0.5	76.5 ± 0.3	93.8 ± 0.5	50.4 ± 1.2	73.5 ± 0.7	69.0 ± 0.5
	SAFT	66.1 ± 0.1	67.2 ± 0.3	27.4 ± 0.8	68.9 ± 3.5	64.8 ± 0.7	84.0 ± 0.4	89.9 ± 0.9	80.7 ± 0.9	93.9 ± 0.2	52.2 ± 1.0	73.7 ± 0.4	69.9 ± 0.5
	CoOp	64.6 ± 0.8	64.7 ± 0.5	27.3 ± 0.1	56.2 ± 6.7	65.4 ± 0.5	83.0 ± 0.6	90.1 ± 1.2	77.0 ± 1.3	92.7 ± 0.7	45.8 ± 0.3	71.2 ± 0.6	67.1 ± 0.5
	KgCoOp	69.2 ± 0.1	68.2 ± 0.2	26.6 ± 0.3	59.4 ± 4.9	65.9 ± 0.4	86.2 ± 0.1	91.4 ± 0.5	72.5 ± 0.2	94.0 ± 0.2	49.0 ± 1.6	71.8 ± 0.3	68.6 ± 0.4
	ProGrad	65.3 ± 0.9	64.2 ± 1.4	27.2 ± 0.9	58.5 ± 6.5	64.4 ± 1.9	84.3 ± 0.1	89.2 ± 0.4	73.0 ± 0.5	93.3 ± 0.2	47.1 ± 1.0	68.9 ± 0.6	66.9 ± 1.0
	TIP-Adapter	67.1 ± 0.1	63.0 ± 0.0	24.2 ± 0.1	45.6 ± 1.6	66.2 ± 0.0	85.3 ± 0.0	88.3 ± 0.1	68.8 ± 0.2	93.4 ± 0.0	44.4 ± 0.0	65.4 ± 0.0	64.7 ± 0.1
	GDA	67.0 ± 0.2	37.6 ± 0.6	20.8 ± 0.8	48.3 ± 5.9	33.4 ± 1.3	43.7 ± 1.4	39.2 ± 2.0	67.4 ± 0.7	77.9 ± 1.9	31.5 ± 2.9	49.9 ± 1.0	47.0 ± 0.5
	ProKeR	67.7 ± 0.1	65.5 ± 0.1	27.5 ± 0.0	59.8 ± 2.0	68.0 ± 0.2	85.4 ± 0.1	88.9 ± 0.2	75.2 ± 0.8	93.7 ± 0.4	49.1 ± 0.7	69.5 ± 0.7	68.2 ± 0.3
	LoRA	63.8 ± 0.5	67.0 ± 0.1	28.8 ± 0.3	65.6 ± 1.8	67.8 ± 0.6	84.8 ± 0.9	91.9 ± 0.7	78.2 ± 0.8	94.1 ± 0.1	52.0 ± 1.0	73.4 ± 0.6	69.8 ± 0.2
	DoRA	63.6 ± 0.4	67.5 ± 0.1	29.0 ± 0.2	66.9 ± 1.6	67.8 ± 0.7	85.0 ± 0.8	92.0 ± 0.7	80.3 ± 1.0	94.1 ± 0.1	52.6 ± 1.1	73.8 ± 0.7	70.2 ± 0.2
	PiSSA	62.5 ± 0.4	68.4 ± 0.2	25.9 ± 0.7	69.6 ± 0.3	66.3 ± 0.9	83.1 ± 0.7	89.5 ± 0.2	82.3 ± 1.1	93.7 ± 0.3	51.8 ± 1.6	73.8 ± 0.5	69.7 ± 0.1
	SO	66.1 ± 0.3	67.9 ± 0.1	30.2 ± 0.2	65.3 ± 2.9	68.3 ± 0.9	86.1 ± 0.5	92.7 ± 0.1	82.2 ± 0.7	94.1 ± 0.2	52.3 ± 0.6	74.1 ± 0.5	70.8 ± 0.2
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.3	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0 ± 0.3
2	ProLIP	68.9 ± 0.1	69.7 ± 0.1	31.0 ± 0.8	74.7 ± 4.4	69.3 ± 0.3	85.4 ± 0.2	90.8 ± 0.2	83.1 ± 1.0	94.8 ± 0.1	58.2 ± 0.7	75.6 ± 0.5	72.9 ± 0.3
	Adapter	66.6 ± 0.3	68.3 ± 0.1	28.8 ± 0.3	71.8 ± 4.4	68.3 ± 0.4	84.9 ± 0.1	91.0 ± 0.6	83.3 ± 0.8	94.8 ± 0.4	57.3 ± 0.6	75.4 ± 0.5	71.9 ± 0.3
	SAFT	66.9 ± 0.2	68.8 ± 0.3	31.1 ± 0.9	78.7 ± 2.6	67.8 ± 0.2	84.8 ± 0.2	92.5 ± 0.6	85.0 ± 0.5	94.6 ± 0.1	57.4 ± 2.1	76.0 ± 0.3	73.1 ± 0.5
	CoOp	67.1 ± 0.5	66.2 ± 0.1	29.9 ± 0.6	69.9 ± 5.8	66.8 ± 0.9	83.3 ± 0.1	91.1 ± 1.8	82.3 ± 1.7	94.1 ± 0.3	54.3 ± 1.2	74.3 ± 0.4	70.8 ± 0.5
	KgCoOp	69.5 ± 0.1	69.3 ± 0.3	27.9 ± 0.3	66.2 ± 2.6	66.4 ± 0.3	86.2 ± 0.4	92.0 ± 0.3	73.6 ± 0.6	94.0 ± 0.1	52.6 ± 0.7	73.3 ± 0.3	70.1 ± 0.2
	ProGrad	66.9 ± 0.9	67.7 ± 0.2	27.0 ± 0.8	67.9 ± 6.2	63.4 ± 5.7	84.4 ± 0.6	89.5 ± 0.2	78.8 ± 1.4	93.2 ± 0.2	51.6 ± 1.3	72.2 ± 1.3	69.3 ± 0.8
	TIP-Adapter	67.2 ± 0.1	63.5 ± 0.1	26.2 ± 0.4	50.2 ± 0.7	67.6 ± 0.2	85.4 ± 0.0	88.4 ± 0.1	75.2 ± 0.1	93.4 ± 0.1	44.5 ± 0.1	65.9 ± 0.1	66.1 ± 0.1
	GDA	67.8 ± 0.1	66.4 ± 0.7	29.6 ± 1.0	70.6 ± 6.6	68.9 ± 0.6	85.4 ± 0.1	88.4 ± 0.1	88.0 ± 0.7	93.9 ± 0.7	54.0 ± 1.0	71.9 ± 0.6	71.4 ± 0.7
	ProKeR	68.4 ± 0.1	67.3 ± 0.2	29.0 ± 0.1	68.4 ± 3.6	70.0 ± 0.7	85.6 ± 0.1	88.4 ± 0.4	78.8 ± 0.3	94.3 ± 0.3	52.7 ± 0.6	71.3 ± 0.3	70.4 ± 0.4
	LoRA	64.7 ± 0.5	68.8 ± 0.3	32.3 ± 0.9	76.4 ± 0.5	67.5 ± 0.7	84.8 ± 0.2	92.5 ± 0.3	85.2 ± 0.3	94.5 ± 0.5	57.1 ± 0.8	76.3 ± 0.1	72.7 ± 0.1
	DoRA	64.5 ± 0.5	69.3 ± 0.3	32.4 ± 1.1	76.5 ± 1.1	67.8 ± 0.6	85.1 ± 0.2	92.6 ± 0.4	85.9 ± 0.1	94.6 ± 0.4	57.4 ± 1.1	76.5 ± 0.2	73.0 ± 0.2
	PiSSA	63.3 ± 0.2	70.1 ± 0.1	30.4 ± 1.4	76.2 ± 3.2	69.5 ± 0.5	84.6 ± 0.2	91.8 ± 0.4	87.0 ± 0.6	94.6 ± 0.1	57.6 ± 0.9	76.8 ± 0.6	72.9 ± 0.4
	SO	65.2 ± 0.3	69.7 ± 0.1	31.1 ± 1.2	76.2 ± 0.9	68.9 ± 0.3	85.8 ± 0.1	92.8 ± 0.7	88.2 ± 0.5	94.9 ± 0.0	58.7 ± 1.2	77.4 ± 0.5	73.5 ± 0.1
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.2 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1 ± 0.2
4	ProLIP	69.6 ± 0.1	71.2 ± 0.2	35.0 ± 0.7	77.7 ± 1.3	72.3 ± 0.2	85.7 ± 0.1	92.0 ± 0.6	88.1 ± 1.1	95.2 ± 0.4	62.5 ± 0.2	78.1 ± 0.2	75.2 ± 0.2
	Adapter	67.4 ± 0.1	69.9 ± 0.2	34.0 ± 0.1	74.5 ± 2.0	71.6 ± 0.3	85.1 ± 0.2	91.5 ± 0.6	90.0 ± 0.6	95.2 ± 0.3	62.3 ± 0.6	79.2 ± 0.4	74.6 ± 0.2
	SAFT	68.2 ± 0.2	69.4 ± 0.2	32.8 ± 0.9	83.6 ± 1.3	70.0 ± 0.1	85.5 ± 0.2	93.2 ± 0.4	88.7 ± 0.3	95.2 ± 0.3	61.6 ± 0.8	77.9 ± 0.9	75.1 ± 0.1
	CoOp	68.4 ± 0.1	68.6 ± 0.3	32.5 ± 0.3	76.3 ± 1.0	70.9 ± 0.2	84.1 ± 0.2	91.7 ± 0.8	87.8 ± 0.4	94.4 ± 0.6	56.7 ± 0.4	75.9 ± 0.7	73.4 ± 0.3
	KgCoOp	69.8 ± 0.0	70.0 ± 0.1	28.8 ± 0.6	64.8 ± 0.3	67.3 ± 0.3	86.5 ± 0.1	92.4 ± 0.2	77.0 ± 0.4	94.6 ± 0.2	54.1 ± 0.7	73.8 ± 0.6	70.8 ± 0.1
	ProGrad	68.2 ± 0.6	69.7 ± 0.4	31.5 ± 1.0	73.7 ± 1.5	70.5 ± 0.3	85.1 ± 0.5	90.5 ± 1.0	82.1 ± 1.2	93.8 ± 0.3	54.7 ± 0.9	73.8 ± 0.2	72.1 ± 0.0
	TIP-Adapter	67.0 ± 0.2	64.4 ± 0.2	30.8 ± 0.3	54.6 ± 0.7	69.4 ± 0.2	85.6 ± 0.0	89.0 ± 0.1	86.8 ± 0.5	93.8 ± 0.1	45.3 ± 0.1	67.1 ± 0.3	68.5 ± 0.0
	GDA	69.5 ± 0.1	71.0 ± 0.1	35.3 ± 0.2	77.7 ± 1.1	71.4 ± 0.4	85.8 ± 0.1	88.6 ± 0.4	93.9 ± 0.6	94.8 ± 0.1	61.2 ± 0.6	76.8 ± 0.4	75.1 ± 0.2
	ProKeR	69.6 ± 0.2	69.1 ± 0.2	31.3 ± 0.3	72.8 ± 1.3	72.5 ± 0.2	86.2 ± 0.1	88.9 ± 0.1	80.6 ± 0.3	94.8 ± 0.2	55.8 ± 0.4	73.6 ± 0.4	72.3 ± 0.0
	LoRA	66.1 ± 0.3	70.2 ± 0.2	33.7 ± 0.5	78.2 ± 2.4	70.7 ± 0.3	85.4 ± 0.2	93.2 ± 0.2	89.0 ± 0.5	95.2 ± 0.2	61.1 ± 0.5	78.4 ± 0.9	74.7 ± 0.3
	DoRA	66.0 ± 0.4	70.6 ± 0.1	33.6 ± 0.3	81.9 ± 1.9	71.0 ± 0.2	85.8 ± 0.2	93.4 ± 0.1	89.6 ± 0.4	95.3 ± 0.2	62.0 ± 0.9	78.6 ± 0.7	75.2 ± 0.2
	PiSSA	64.8 ± 0.3	71.5 ± 0.1	34.4 ± 0.4	83.7 ± 1.3	72.7 ± 0.2	86.0 ± 0.2	93.3 ± 0.1	91.2 ± 0.2	95.4 ± 0.4	62.8 ± 0.7	79.8 ± 0.4	76.0 ± 0.2
	SO	64.1 ± 0.1	71.4 ± 0.1	34.6 ± 0.1	81.0 ± 1.6	71.9 ± 0.7	86.2 ± 0.1	93.6 ± 0.1	93.0 ± 0.1	95.5 ± 0.5	64.1 ± 0.5	79.8 ± 0.6	75.9 ± 0.1
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.0	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8 ± 0.1
8	ProLIP	70.2 ± 0.2	72.6 ± 0.4	38.7 ± 0.5	82.1 ± 0.8	74.2 ± 0.4	85.8 ± 0.4	92.3 ± 0.3	90.7 ± 1.6	95.1 ± 0.3	66.3 ± 0.8	81.0 ± 0.3	77.2 ± 0.0
	Adapter	68.5 ± 0.3	71.8 ± 0.3	36.3 ± 0.9	80.4 ± 1.8	75.3 ± 0.3	85.2 ± 0.1	91.8 ± 0.4	93.6 ± 0.6	95.7 ± 0.1	66.8 ± 0.9	82.5 ± 0.4	77.1 ± 0.2
	SAFT	69.6 ± 0.3	70.7 ± 0.2	36.3 ± 0.9	88.9 ± 0.9	73.5 ± 0.5	85.7 ± 0.3	93.3 ± 0.1	91.8 ± 0.3	95.8 ± 0.1	65.4 ± 1.4	79.6 ± 0.5	77.3 ± 0.3
	CoOp	70.2 ± 0.3	70.9 ± 0.2	36.3 ± 0.8	77.5 ± 1.5	74.4 ± 0.6	83.9 ± 0.5	91.7 ± 0.3	92.5 ± 0.2	95.2 ± 0.5	61.6 ± 0.8	78.7 ± 0.8	75.7 ± 0.2
	KgCoOp	70.1 ± 0.0	70.9 ± 0.4	31.1 ± 0.4	70.5 ± 0.7	68.2 ± 0.1	86.8 ± 0.2	92.8 ± 0.1	81.4 ± 0.1	94.8 ± 0.1	58.6 ± 0.2	75.8 ± 0.3	72.8 ± 0.0
	ProGrad	69.7 ± 0.5	71.4 ± 0.3	35.1 ± 0.7	76.2 ± 2.0	73.4 ± 0.6	85.4 ± 0.2	91.5 ± 0.4	86.8 ± 0.6	94.7 ± 0.1	58.9 ± 0.3	75.2 ± 0.9	74.4 ± 0.2
	TIP-Adapter	67.0 ± 0.4	66.3 ± 0.1	34.7 ± 0.6	70.4 ± 2.4	72.5 ± 0.2	86.0 ± 0.0	89.7 ± 0.1	92.5 ± 0.5	94.1 ± 0.1	49.1 ± 0.2	69.6 ± 0.2	72.0 ± 0.3
	GDA	71.0 ± 0.2	72.6 ± 0.7	41.4 ± 1.0	83.2 ± 1.0	76.9 ± 0.3	85.9 ± 0.1	88.7 ± 1.6	96.0 ± 0.3	94.9 ± 0.4	65.9 ± 1.1	81.1 ± 0.5	78.0 ± 0.1
	ProKeR	70.1 ± 0.2	72.1 ± 0.1	35.8 ± 0.5	81.3 ± 0.9	76.5 ± 0.1	86.7 ± 0.0	89.8 ± 0.3	84.6 ± 0.3	95.5 ± 0.1	61.2 ± 0.6	77.7 ± 0.6	75.6 ± 0.1
	LoRA	68.0 ± 0.2	72.2 ± 0.3	39.0 ± 0.5	86.7 ± 1.3	75.7 ± 0.3	85.5 ± 0.2	93.3 ± 0.2	92.7 ± 0.5	95.8 ± 0.3	65.5 ± 1.3	80.3 ± 0.5	77.7 ± 0.3
	DoRA	68.0 ± 0.3	72.4 ± 0.3	39.4 ± 0.2	88.0 ± 0.8	76.0 ± 0.5	85.8 ± 0.2	93.5 ± 0.2	92.9 ± 0.2	95.9 ± 0.3	66.2 ± 1.2	80.6 ± 0.6	78.1 ± 0.2
	PiSSA	66.6 ± 0.1	72.8 ± 0.2	39.3 ± 0.5	89.1 ± 1.1	76.5 ± 0.2	86.1 ± 0.2	93.4 ± 0.2	93.8 ± 0.1	95.8 ± 0.4	67.1 ± 1.2	81.6 ± 0.3	78.4 ± 0.2
	SO	62.8 ± 0.3	72.8 ± 0.2	39.5 ± 0.2	88.9 ± 0.8	76.9 ± 0.3	86.2 ± 0.1	93.2 ± 0.4	95.3 ± 0.2	96.4 ± 0.2	69.2 ± 0.7	81.4 ± 0.2	78.4 ± 0.0
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.3	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3 ± 0.1
16	ProLIP	71.2 ± 0.1	74.1 ± 0.1	43.8 ± 1.3	87.2 ± 0.6	77.4 ± 0.6	85.7 ± 0.8	92.7 ± 0.1	94.9 ± 1.2	95.7 ± 0.1	70.5 ± 0.1	83.0 ± 0.6	79.7 ± 0.3
	Adapter	69.9 ± 0.1	73.7 ± 0.3	41.9 ± 1.1	86.1 ± 0.4	78.9 ± 0.5	85.4 ± 0.3	92.1 ± 0.4	96.4 ± 0.2	95.9 ± 0.3	70.5 ± 0.5	83.8 ± 0.2	79.5 ± 0.1
	SAFT	71.4 ± 0.1	72.2 ± 0.1	41.8 ± 0.2	93.1 ± 0.5	77.1 ± 0.1	85.8 ± 0.3	93.6 ± 0.4	95.1 ± 0.1	95.9 ± 0.1	69.6 ± 0.7	82.5 ± 0.2	79.8 ± 0.1
	CoOp	71.8 ± 0.2	73.5 ± 0.2	39.5 ± 0.2	84.4 ± 1.0	78.6 ± 0.4	84.5 ± 0.4	92.4 ± 0.1	95.4 ± 0.1	95.3 ± 0.0	66.4 ± 0.2	81.0 ± 0.1	78.4 ± 0.1
	KgCoOp	70.3 ± 0.0	72.0 ± 0.2	32.9 ± 0.2	73.0 ± 0.9	69.4 ± 0.3	86.8 ± 0.0	93.1 ± 0.1	86.7 ± 0.2	95.0 ± 0.2	63.8 ± 0.2	77.5 ± 0.2	74.6 ± 0.1
	ProGrad	61.7 ± 7.9	73.5 ± 0.2	39.0 ± 0.7	79.7 ± 2.8	76.2 ± 0.2	85.6 ± 0.3	91.6 ± 0.1	89.9 ± 0.5	94.8 ± 0.4	62.6 ± 0.9	77.9 ± 0.8	75.7 ± 0.4
	TIP-Adapter	68.3 ± 0.1	69.9 ± 0.1	41.1 ± 0.1	82.8 ± 0.1	76.8 ± 0.3	86.6 ± 0.1	91.6 ± 0.5	95.2 ± 0.6	94.6 ± 0.1	55.1 ± 0.3	73.2 ± 0.2	75.9 ± 0.1
	GDA	71.3 ± 0.1	75.8 ± 0.0	48.3 ± 0.5	89.1 ± 0.3	81.8 ± 0.3	86.9 ± 0.0	91.6 ± 1.1	97.5 ± 0.4	95.6 ± 0.2	70.8 ± 0.1	83.4 ± 0.2	81.1 ± 0.1
	ProKeR	71.9 ± 0.1	75.4 ± 0.1	42.9 ± 0.5	87.3 ± 0.7	81.3 ± 0.3	87.3 ± 0.1	91.6 ± 0.3	89.7 ± 0.0	96.3 ± 0.1	68.9 ± 0.5	81.7 ± 0.3	79.5 ± 0.0
	LoRA	70.5 ± 0.2	73.8 ± 0.1	46.1 ± 0.7	92.0 ± 0.9	79.7 ± 0.1	85.8 ± 0.3	93.6 ± 0.3	95.8 ± 0.2	95.9 ± 0.0	70.5 ± 0.3	82.9 ± 0.5	80.6 ± 0.0
	DoRA	70.4 ± 0.2	73.9 ± 0.1	46.4 ± 0.6	92.3 ± 0.8	80.1 ± 0.3	86.1 ± 0.3	93.8 ± 0.2	95.6 ± 0.2	95.9 ± 0.1	70.7 ± 0.1	83.1 ± 0.6	80.7 ± 0.0
	PiSSA	69.0 ± 0.2	74.2 ± 0.2	45.8 ± 0.2	93.1 ± 0.7	79.9 ± 0.3	86.5 ± 0.2	93.9 ± 0.2	96.4 ± 0.0	96.3 ± 0.1	71.4 ± 0.3	83.8 ± 0.5	80.9 ± 0.1
	SO	60.5 ± 0.2	74.0 ± 0.1	46.0 ± 0.4	92.9 ± 0.8	80.2 ± 0.0	86.4 ± 0.1	93.6 ± 0.3	97.5 ± 0.1	96.6 ± 0.3	72.9 ± 0.4	84.1 ± 0.2	80.4 ± 0.0
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.0	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6 ± 0.0

Table 22: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	ProLIP	75.4 ± 0.1	71.8 ± 0.1	38.0 ± 0.6	71.4 ± 1.8	79.7 ± 0.3	90.4 ± 0.1	94.0 ± 0.2	87.4 ± 0.3	96.1 ± 0.3	58.5 ± 0.2	80.4 ± 0.5	76.6 ± 0.1
	Adapter	72.5 ± 0.3	71.4 ± 0.3	37.7 ± 0.8	70.0 ± 2.7	77.9 ± 0.4	90.1 ± 0.2	92.7 ± 0.6	91.0 ± 0.1	96.3 ± 0.6	58.7 ± 0.9	81.0 ± 0.6	76.3 ± 0.1
	SAFT	74.2 ± 0.6	73.7 ± 0.1	38.4 ± 1.6	74.2 ± 3.5	78.5 ± 0.0	89.4 ± 0.4	91.0 ± 1.9	89.1 ± 0.8	96.2 ± 0.8	59.6 ± 1.7	81.3 ± 0.8	76.9 ± 0.7
	CoOp	71.2 ± 0.6	67.1 ± 0.5	35.6 ± 0.5	65.7 ± 3.0	77.1 ± 0.6	89.1 ± 0.5	93.3 ± 0.6	89.0 ± 0.7	94.9 ± 0.2	53.3 ± 0.8	77.5 ± 0.6	74.0 ± 0.2
	KgCoOp	75.8 ± 0.0	71.8 ± 0.4	35.2 ± 0.9	70.4 ± 1.6	78.1 ± 0.5	90.9 ± 0.2	94.3 ± 0.2	82.0 ± 0.2	96.3 ± 0.3	57.0 ± 0.9	78.1 ± 0.5	75.4 ± 0.1
	ProGrad	70.9 ± 0.6	68.2 ± 0.7	34.1 ± 2.0	63.3 ± 3.6	77.4 ± 0.9	90.0 ± 0.1	93.1 ± 1.6	83.6 ± 0.9	93.8 ± 1.3	53.1 ± 1.2	74.7 ± 0.1	72.9 ± 0.4
	TIP-Adapter	73.6 ± 0.0	68.0 ± 0.0	31.0 ± 0.3	56.7 ± 0.4	77.6 ± 0.1	90.3 ± 0.0	93.1 ± 0.1	78.4 ± 0.1	95.2 ± 0.0	52.4 ± 0.0	73.9 ± 0.0	71.8 ± 0.1
	GDA	73.9 ± 0.2	40.5 ± 0.6	27.4 ± 0.2	52.4 ± 4.5	45.7 ± 1.3	55.1 ± 1.7	47.6 ± 0.9	83.2 ± 0.9	80.5 ± 1.5	34.8 ± 1.8	57.4 ± 1.7	54.4 ± 0.4
	ProKeR	74.4 ± 0.1	69.9 ± 0.2	35.9 ± 0.3	66.3 ± 1.2	78.5 ± 0.3	90.4 ± 0.1	93.5 ± 0.5	83.1 ± 0.8	95.5 ± 0.3	55.4 ± 0.3	76.5 ± 0.1	74.5 ± 0.1
	LoRA	72.4 ± 0.6	73.5 ± 0.3	40.8 ± 1.0	72.9 ± 0.9	79.3 ± 0.2	90.5 ± 0.3	94.2 ± 0.2	87.3 ± 0.4	96.7 ± 0.5	58.5 ± 1.4	81.0 ± 0.4	77.0 ± 0.3
	DoRA	72.2 ± 0.5	73.5 ± 0.2	40.4 ± 1.1	72.6 ± 0.9	79.4 ± 0.2	90.7 ± 0.2	94.4 ± 0.1	88.6 ± 0.7	96.7 ± 0.5	59.4 ± 1.2	81.3 ± 0.6	77.2 ± 0.1
	PiSSA	72.1 ± 0.6	74.3 ± 0.1	38.0 ± 1.4	73.0 ± 4.9	78.8 ± 0.2	89.5 ± 0.1	93.1 ± 0.2	90.6 ± 0.5	96.4 ± 0.3	59.6 ± 2.0	81.5 ± 0.6	77.0 ± 0.6
	SO	74.2 ± 0.1	72.4 ± 0.1	39.6 ± 1.3	72.3 ± 1.0	78.9 ± 0.5	91.3 ± 0.2	94.7 ± 0.1	89.3 ± 0.4	96.6 ± 0.5	59.9 ± 1.1	81.5 ± 0.4	77.3 ± 0.1
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.6 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4 ± 0.1
2	ProLIP	75.9 ± 0.0	73.5 ± 0.1	40.8 ± 0.6	79.2 ± 1.3	80.6 ± 0.2	90.6 ± 0.3	94.2 ± 0.3	90.7 ± 0.2	96.5 ± 0.3	61.5 ± 0.6	81.6 ± 0.5	78.6 ± 0.1
	Adapter	73.2 ± 0.3	72.9 ± 0.3	40.0 ± 1.0	79.1 ± 1.4	79.2 ± 0.6	90.2 ± 0.1	93.4 ± 0.6	94.1 ± 0.4	96.7 ± 0.2	62.4 ± 0.4	82.6 ± 0.3	78.5 ± 0.2
	SAFT	74.7 ± 0.3	75.1 ± 0.2	40.9 ± 1.4	85.1 ± 2.6	80.0 ± 0.2	90.7 ± 0.2	94.2 ± 0.2	92.7 ± 0.7	97.1 ± 0.3	66.8 ± 0.9	82.4 ± 0.8	80.0 ± 0.2
	CoOp	73.3 ± 0.2	70.1 ± 0.6	39.3 ± 1.4	77.8 ± 1.6	77.5 ± 0.4	89.0 ± 0.7	93.4 ± 0.8	92.3 ± 0.9	96.0 ± 0.4	58.7 ± 0.4	80.5 ± 0.7	77.1 ± 0.3
	KgCoOp	76.1 ± 0.1	72.9 ± 0.0	36.9 ± 0.7	75.3 ± 2.1	78.4 ± 0.5	91.0 ± 0.2	94.4 ± 0.4	84.9 ± 1.3	96.6 ± 0.1	60.2 ± 0.3	79.5 ± 0.3	76.9 ± 0.3
	ProGrad	68.3 ± 7.6	70.5 ± 0.4	31.7 ± 7.5	72.1 ± 3.8	78.7 ± 0.2	89.8 ± 0.7	93.9 ± 0.4	86.9 ± 1.3	95.6 ± 0.1	55.5 ± 0.9	77.4 ± 1.1	74.6 ± 0.7
	TIP-Adapter	73.7 ± 0.2	68.3 ± 0.1	34.4 ± 0.7	58.5 ± 0.3	78.6 ± 0.3	90.4 ± 0.0	93.3 ± 0.0	84.3 ± 0.5	95.4 ± 0.1	52.7 ± 0.0	74.4 ± 0.2	73.1 ± 0.1
	GDA	75.0 ± 0.0	71.3 ± 0.1	39.3 ± 0.9	78.3 ± 0.7	80.2 ± 0.2	90.1 ± 0.2	92.9 ± 0.4	94.5 ± 2.0	96.1 ± 0.6	60.2 ± 3.5	79.7 ± 0.3	78.0 ± 0.3
	ProKeR	75.0 ± 0.0	71.4 ± 0.2	37.9 ± 0.7	73.6 ± 0.6	79.8 ± 0.6	90.7 ± 0.0	93.7 ± 0.5	84.9 ± 0.8	96.3 ± 0.2	58.0 ± 0.6	77.7 ± 0.2	76.3 ± 0.0
	LoRA	72.9 ± 0.3	74.4 ± 0.2	42.7 ± 2.0	77.7 ± 1.7	79.9 ± 0.2	90.8 ± 0.3	94.7 ± 0.3	92.5 ± 0.6	97.2 ± 0.2	63.8 ± 0.8	82.1 ± 1.0	79.0 ± 0.2
	DoRA	72.8 ± 0.3	74.6 ± 0.2	42.1 ± 1.8	80.1 ± 1.6	80.1 ± 0.2	90.9 ± 0.2	94.8 ± 0.2	92.7 ± 0.7	97.2 ± 0.2	64.7 ± 0.7	82.1 ± 0.7	79.3 ± 0.2
	PiSSA	72.2 ± 0.2	75.5 ± 0.1	41.2 ± 1.8	84.7 ± 2.0	80.9 ± 0.2	90.4 ± 0.4	94.3 ± 0.3	93.9 ± 0.6	97.5 ± 0.3	66.2 ± 1.4	83.0 ± 0.9	80.0 ± 0.3
	SO	73.6 ± 0.2	70.3 ± 0.3	42.9 ± 1.6	80.4 ± 1.6	79.7 ± 0.1	91.0 ± 0.3	95.1 ± 0.1	94.1 ± 0.3	97.3 ± 0.3	66.0 ± 0.9	82.9 ± 0.4	79.4 ± 0.3
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1 ± 0.1
4	ProLIP	76.6 ± 0.1	75.0 ± 0.1	45.2 ± 0.4	82.9 ± 1.9	82.3 ± 0.5	90.8 ± 0.1	94.5 ± 0.2	93.8 ± 0.1	97.3 ± 0.2	66.2 ± 0.5	84.0 ± 0.3	80.8 ± 0.2
	Adapter	74.0 ± 0.2	74.7 ± 0.2	44.9 ± 0.7	83.3 ± 1.5	81.6 ± 0.4	90.5 ± 0.0	94.1 ± 0.3	96.7 ± 0.2	97.4 ± 0.3	67.2 ± 0.4	85.2 ± 0.7	80.9 ± 0.3
	SAFT	75.7 ± 0.3	75.8 ± 0.1	44.9 ± 0.7	87.7 ± 1.9	82.3 ± 0.5	91.1 ± 0.1	95.2 ± 0.2	95.5 ± 0.3	97.4 ± 0.3	68.4 ± 0.4	85.2 ± 0.5	81.7 ± 0.1
	CoOp	74.6 ± 0.1	72.4 ± 0.2	42.3 ± 0.4	81.9 ± 1.5	80.6 ± 0.5	89.7 ± 0.2	94.6 ± 0.5	95.2 ± 0.4	96.9 ± 0.2	63.2 ± 0.5	81.4 ± 0.3	79.3 ± 0.1
	KgCoOp	76.4 ± 0.1	74.0 ± 0.1	39.0 ± 0.5	76.5 ± 2.5	79.0 ± 0.2	91.2 ± 0.1	94.7 ± 0.1	86.3 ± 0.2	96.8 ± 0.4	62.7 ± 0.6	80.6 ± 0.4	77.9 ± 0.3
	ProGrad	74.9 ± 0.3	72.5 ± 0.0	28.6 ± 18.4	77.6 ± 3.6	80.6 ± 0.5	90.6 ± 0.1	87.3 ± 9.8	88.4 ± 0.3	95.9 ± 0.3	59.9 ± 0.2	78.6 ± 0.5	75.9 ± 1.8
	TIP-Adapter	73.5 ± 0.1	68.9 ± 0.1	39.6 ± 0.5	60.7 ± 0.1	79.5 ± 0.2	90.5 ± 0.0	93.6 ± 0.1	92.1 ± 0.3	95.6 ± 0.0	53.1 ± 0.1	75.0 ± 0.1	74.7 ± 0.1
	GDA	76.9 ± 0.1	73.7 ± 0.6	46.6 ± 0.6	83.0 ± 0.9	82.5 ± 0.3	90.5 ± 0.2	93.6 ± 0.6	97.9 ± 0.1	97.1 ± 0.2	67.8 ± 1.2	83.5 ± 1.5	81.2 ± 0.1
	ProKeR	76.1 ± 0.1	73.1 ± 0.0	40.9 ± 0.1	75.7 ± 1.5	82.0 ± 0.1	91.1 ± 0.0	94.3 ± 0.2	87.3 ± 0.1	96.8 ± 0.2	61.8 ± 0.3	79.8 ± 0.4	78.1 ± 0.1
	LoRA	73.7 ± 0.3	75.8 ± 0.2	46.2 ± 1.0	82.9 ± 0.6	82.6 ± 0.2	90.9 ± 0.2	95.4 ± 0.1	95.6 ± 0.1	97.4 ± 0.4	67.5 ± 0.2	85.2 ± 0.5	81.2 ± 0.0
	DoRA	73.7 ± 0.2	76.0 ± 0.2	45.9 ± 0.6	83.7 ± 2.2	82.6 ± 0.4	91.0 ± 0.2	95.4 ± 0.2	95.6 ± 0.1	97.4 ± 0.4	67.6 ± 0.0	85.2 ± 0.4	81.3 ± 0.2
	PiSSA	73.2 ± 0.1	76.9 ± 0.2	47.1 ± 0.4	87.3 ± 1.6	83.3 ± 0.3	91.1 ± 0.1	95.2 ± 0.1	96.4 ± 0.3	97.5 ± 0.4	69.2 ± 0.4	86.2 ± 0.2	82.1 ± 0.1
	SO	73.6 ± 0.2	71.5 ± 0.2	47.3 ± 0.7	82.8 ± 2.8	78.5 ± 0.6	90.6 ± 0.1	95.2 ± 0.0	95.9 ± 0.3	97.5 ± 0.3	69.9 ± 0.2	86.1 ± 0.3	80.8 ± 0.2
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.3	91.3 ± 0.2	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.4	68.7 ± 0.9	84.9 ± 0.5	82.5 ± 0.2
8	ProLIP	77.3 ± 0.1	76.7 ± 0.1	49.7 ± 0.8	87.1 ± 0.9	84.3 ± 0.2	91.2 ± 0.0	94.7 ± 0.4	95.4 ± 0.2	97.4 ± 0.2	69.6 ± 0.7	85.4 ± 0.1	82.6 ± 0.2
	Adapter	75.0 ± 0.2	76.2 ± 0.2	50.0 ± 0.7	87.0 ± 1.7	84.2 ± 0.5	90.7 ± 0.0	94.0 ± 0.7	97.7 ± 0.2	97.6 ± 0.3	71.2 ± 0.6	86.6 ± 0.1	82.7 ± 0.3
	SAFT	76.6 ± 0.1	76.8 ± 0.2	51.4 ± 0.6	90.3 ± 1.6	84.7 ± 0.4	91.5 ± 0.2	95.6 ± 0.2	96.9 ± 0.1	97.6 ± 0.2	72.1 ± 0.3	86.0 ± 0.3	83.6 ± 0.2
	CoOp	76.2 ± 0.1	74.6 ± 0.5	48.3 ± 1.0	84.1 ± 1.3	83.4 ± 0.2	89.9 ± 0.2	94.3 ± 0.5	96.9 ± 0.4	96.8 ± 0.2	67.3 ± 1.4	84.1 ± 1.0	81.5 ± 0.4
	KgCoOp	76.6 ± 0.1	74.8 ± 0.2	40.6 ± 0.4	80.3 ± 1.2	79.7 ± 0.3	91.4 ± 0.0	94.5 ± 0.2	89.3 ± 1.3	97.0 ± 0.2	65.7 ± 0.3	81.8 ± 0.6	79.3 ± 0.2
	ProGrad	75.7 ± 0.8	74.7 ± 0.3	46.5 ± 1.2	79.0 ± 1.9	82.4 ± 0.4	90.5 ± 0.2	94.5 ± 0.1	91.7 ± 0.4	96.0 ± 0.5	63.2 ± 0.9	80.4 ± 0.2	79.5 ± 0.3
	TIP-Adapter	73.6 ± 0.1	70.4 ± 0.1	43.7 ± 0.5	70.1 ± 1.6	81.9 ± 0.3	90.8 ± 0.1	94.1 ± 0.2	95.7 ± 0.4	95.7 ± 0.2	55.0 ± 0.4	76.0 ± 0.1	77.0 ± 0.2
	GDA	77.3 ± 0.2	77.4 ± 0.1	53.1 ± 1.1	87.1 ± 0.4	85.8 ± 0.5	91.3 ± 0.1	94.1 ± 0.8	98.6 ± 0.1	97.5 ± 0.1	70.3 ± 0.3	85.4 ± 0.3	83.4 ± 0.1
	ProKeR	76.5 ± 0.2	75.8 ± 0.1	46.1 ± 0.6	83.8 ± 0.5	85.0 ± 0.2	91.4 ± 0.1	95.1 ± 0.3	90.8 ± 0.1	97.4 ± 0.2	65.9 ± 0.8	82.4 ± 0.3	80.9 ± 0.0
	LoRA	75.3 ± 0.3	77.4 ± 0.3	52.9 ± 0.8	89.2 ± 1.2	85.6 ± 0.3	91.2 ± 0.3	95.6 ± 0.1	97.0 ± 0.2	97.7 ± 0.3	72.3 ± 0.3	86.0 ± 0.5	83.6 ± 0.2
	DoRA	75.2 ± 0.3	77.6 ± 0.3	53.6 ± 1.0	90.0 ± 1.2	85.8 ± 0.4	91.3 ± 0.3	95.7 ± 0.1	97.0 ± 0.1	97.7 ± 0.3	72.3 ± 0.2	86.1 ± 0.5	83.8 ± 0.2
	PiSSA	74.0 ± 0.1	78.1 ± 0.2	53.4 ± 0.6	91.0 ± 1.1	85.8 ± 0.4	91.5 ± 0.1	95.6 ± 0.2	97.5 ± 0.2	97.8 ± 0.4	73.3 ± 0.3	87.1 ± 0.0	84.1 ± 0.2
	SO	73.4 ± 0.2	73.8 ± 0.4	50.4 ± 1.0	89.7 ± 1.1	83.0 ± 2.3	89.2 ± 0.4	95.2 ± 0.3	96.7 ± 0.5	97.8 ± 0.3	73.6 ± 0.4	85.8 ± 0.3	82.6 ± 0.2
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7 ± 0.2
16	ProLIP	78.2 ± 0.1	78.1 ± 0.1	55.3 ± 0.4	91.1 ± 0.2	85.9 ± 0.1	91.3 ± 0.2	95.0 ± 0.2	97.1 ± 0.2	97.7 ± 0.1	73.9 ± 0.3	87.3 ± 0.1	84.6 ± 0.0
	Adapter	76.0 ± 0.0	77.7 ± 0.2	56.0 ± 0.3	91.3 ± 0.3	86.4 ± 0.3	90.8 ± 0.2	94.2 ± 0.2	98.5 ± 0.1	97.8 ± 0.1	74.8 ± 0.3	87.8 ± 0.3	84.7 ± 0.0
	SAFT	77.6 ± 0.2	77.7 ± 0.2	59.0 ± 0.4	94.1 ± 0.8	86.5 ± 0.1	91.5 ± 0.2	95.8 ± 0.1	97.9 ± 0.3	97.8 ± 0.2	74.9 ± 0.2	87.8 ± 0.2	85.5 ± 0.2
	CoOp	77.8 ± 0.1	76.8 ± 0.1	53.0 ± 1.2	89.7 ± 0.2	86.2 ± 0.4	89.8 ± 0.4	94.1 ± 0.2	98.3 ± 0.2	97.3 ± 0.1	72.0 ± 0.7	85.8 ± 0.1	83.7 ± 0.2
	KgCoOp	76.9 ± 0.1	75.6 ± 0.1	42.6 ± 0.7	84.2 ± 1.3	80.5 ± 0.2	91.4 ± 0.1	94.8 ± 0.1	92.0 ± 0.3	97.4 ± 0.2	69.3 ± 0.7	83.5 ± 0.4	80.7 ± 0.1
	ProGrad	68.7 ± 11.0	76.5 ± 0.1	50.3 ± 0.8	83.0 ± 2.5	84.3 ± 0.3	90.5 ± 0.3	94.7 ± 0.4	93.8 ± 0.4	96.6 ± 0.2	69.1 ± 0.4	82.4 ± 0.1	80.9 ± 0.9
	TIP-Adapter	74.3 ± 0.2	73.3 ± 0.2	50.1 ± 0.3	85.6 ± 0.3	84.8 ± 0.1	91.1 ± 0.0	94.6 ± 0.1	97.3 ± 0.1	96.2 ± 0.2	57.8 ± 0.1	78.3 ± 0.3	80.3 ± 0.1
	GDA	78.5 ± 0.0	79.8 ± 0.1	59.9 ± 0.2	86.6 ± 0.8	87.8 ± 0.4	91.1 ± 0.5	94.7 ± 0.1	99.2 ± 0.1	97.9 ± 0.1	74.7 ± 0.5	87.3 ± 0.3	85.2 ± 0.0
	ProKeR	78.0 ± 0.1	78.7 ± 0.1	54.5 ± 0.1	90.3 ± 0.5	88.0 ± 0.2	91.8 ± 0.0	95.3 ± 0.2	94.6 ± 0.1	98.0 ± 0.1	73.2 ± 0.8	85.6 ± 0.1	84.4 ± 0.1
	LoRA	76.8 ± 0.2	78.7 ± 0.2	61.1 ± 0.5	93.1 ± 0.5	87.8 ± 0.2	91.2 ± 0.2	95.8 ± 0.0	98.3 ± 0.1	97.7 ± 0.1	75.4 ± 0.2	88.0 ± 0.1	85.8 ± 0.1
	DoRA	76.9 ± 0.1	78.9 ± 0.1	61.1 ± 0.6	93.3 ± 0.5	87.8 ± 0.2	91.3 ± 0.2	95.9 ± 0.1	98.3 ± 0.1	97.7 ± 0.1	75.1 ± 0.1	88.1 ± 0.1	85.9 ± 0.1
	PiSSA	75.7 ± 0.0	79.1 ± 0.1	61.7 ± 0.8	94.0 ± 0.5	87.5 ± 0.2	91.6 ± 0.2	95.9 ± 0.1	98.6 ± 0.2	97.8 ± 0.1	76.0 ± 0.7	88.6 ± 0.3	86.0 ± 0.1
	SO	73.4 ± 0.1	76.1 ± 0.5	56.6 ± 0.2	93.6 ± 0.8	85.7 ± 1.2	87.1 ± 0.0	95.4 ± 0.2	98.3 ± 0.1	97.9 ± 0.1	74.7 ± 0.4	85.8 ± 0.1	84.0 ± 0.2
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8 ± 0.1

F.2 Cross-Dataset Transfer of the Selected Training Epoch

Diagnostic and selection rule. Figure 4 evaluates 1-shot CLIP (ViT-B/16) adaptation at $\eta = 10^{-4}$. For each method and source dataset, we first average test accuracy across the three seeds at each epoch. We select the earliest maximizer over epochs 1–300. We then read each target's mean accuracy at exactly this epoch from its independently trained trajectory. Only the training duration is transferred. Neither trained weights nor source examples are reused on the target. Each heatmap entry reports the target accuracy change relative to its own zero-shot CLIP baseline.

The source epoch is selected using *test* accuracy. The diagonal of the matrix provided in Figure 4 contains each dataset's own best seed-mean accuracy relative to zero-shot CLIP. We exclude these 11 diagonal entries when summarizing transfer.

Table 23: Epoch-transfer diagnostic at $\eta = 10^{-4}$ for 1-shot CLIP (ViT-B/16). The target oracle is the dataset-average maximum of the seed-mean accuracy curve. Transfer accuracy and the gap to that oracle average all 110 off-diagonal source–target pairs. The final column counts pairs within one percentage point of the target oracle. Only the source-selected epoch is reused. These are test-guided diagnostics, not validation-selected estimates.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp
LoRA	71.34	68.05	3.29	46/110
SO	72.20	69.78	2.43	55/110
SAFT	71.35	67.57	3.78	45/110
OLM-SGD	**72.64**	**71.73**	**0.91**	**85/110**

Accuracy retained under a transferred duration. Table 23 averages the 110 ordered source–target pairs with distinct datasets. Its oracle gap is the target's best seed-mean accuracy minus the accuracy at the source-selected epoch. OLM-SGD achieves 71.73% mean transfer accuracy and an oracle gap of 0.91 points. The corresponding gap is 2.43 for SO, 3.29 for LoRA, and 3.78 for SAFT. OLM-SGD stays within one point of the target oracle on $85/110$ pairs, compared with $55/110$ for SO and fewer than half for LoRA and SAFT.

This difference is not explained only by higher attainable accuracy. OLM-SGD's target-oracle average exceeds SO's by 0.44 points, whereas its transfer average is 1.96 points higher. The remaining difference is the smaller penalty from transferring the duration. Thus, the diagnostic supports a broader range of useful stopping times rather than merely a better task-specific peak.

ImageNet remains the main exception. OLM-SGD's selected epochs range from 18 on ImageNet to 292 on EuroSAT. Despite this range, $103/110$ off-diagonal transfers improve over zero-shot CLIP in Figure 4. All seven negative entries target ImageNet. The largest displayed ImageNet decrease is 4.1 points for OLM-SGD, compared with 18.7 for LoRA, 18.8 for SO, and 12.9 for SAFT. All 100 off-diagonal transfers to the other ten targets remain positive for OLM-SGD. This pattern supports reduced sensitivity to the source-selected duration, but not universal epoch transfer. In particular, a late epoch selected on another dataset can still damage ImageNet accuracy.

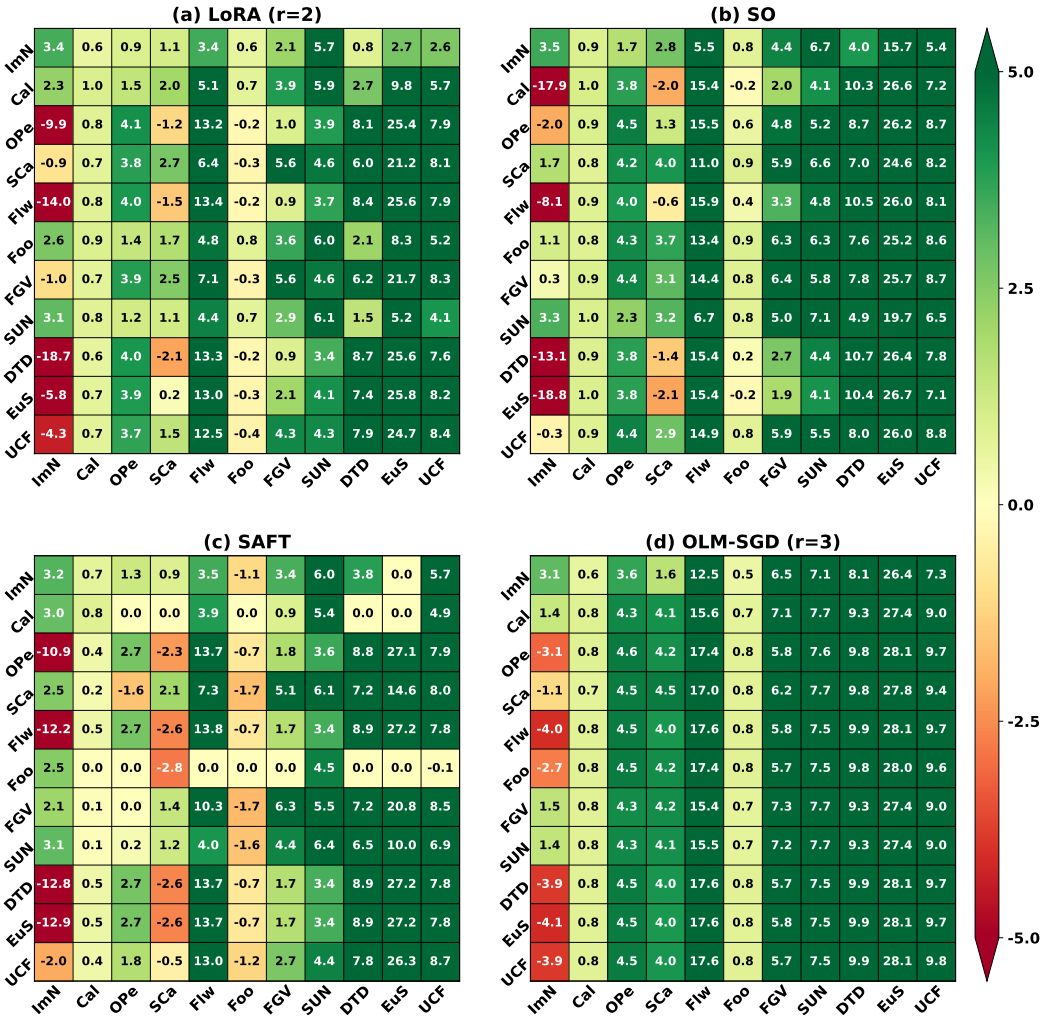

Figure 4: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) adaptation at $\eta = 10^{-4}$. We compare LoRA (rank 2), SO and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). Each row selects the earliest epoch in 1–300 that maximizes the source's mean *test* accuracy across three seeds. Each column evaluates an independently trained target at that epoch. Values are target accuracy changes relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Only the duration is transferred. The diagonal uses target-specific oracle selection. This is a test-guided diagnostic, not validation-selected early stopping.

F.3 Long-Horizon Dynamics and Learning-Rate Sensitivity

Protocol and coverage. Figures 5 and 6 show test accuracy over 300 epochs. Figures 7 and 8 show the corresponding training loss. These runs use 1-shot CLIP (ViT-B/16) on all 11 datasets with learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The four methods retain the settings stated above. All $4 \times 4 \times 11 \times 3 = 528$ trajectories contain epochs 1–300. Curves report seed means and standard deviations. This common horizon complements the method-specific schedules of the few-shot tables.

Retaining accuracy during prolonged training. At 10^{-4}, OLM-SGD reaches 71.75% mean final accuracy. The three baselines range from 67.98% to 68.54% (Table 24). OLM-SGD has the highest final accuracy on nine datasets. SO leads on Caltech101 and DTD. From epoch 30 to epoch 300, OLM-SGD's mean accuracy changes by $+0.14$ points. LoRA, SO, and SAFT instead lose 1.76, 1.92, and 1.43 points. These averages support improved retention over a prolonged horizon.

Learning-rate sensitivity is reduced, not removed. Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of OLM-SGD's mean final accuracy is 2.52 points. The corresponding ranges are 3.79 for LoRA, 18.76 for SO, and 16.38 for SAFT. OLM-SGD reaches almost identical aggregate accuracy at 10^{-4} and 10^{-5}, both rounding to 71.75%. However, the smaller rate also makes SO substantially more competitive. SO reaches 71.23% at 10^{-5} and leads on five of the 11 datasets. Its mean gap to OLM-SGD is only 0.52 points at that rate.

The similar OLM-SGD averages conceal dataset-specific trade-offs. Reducing the rate from 10^{-4} to 10^{-5} raises its final ImageNet accuracy from 62.48% to 69.44%. The same change lowers Flowers102 accuracy from 85.02% to 81.36%. At 10^{-2}, OLM-SGD's mean final accuracy falls to 47.38%. Although this exceeds the three baselines at that rate, it remains well below the reported 64.1% zero-shot average. The evidence therefore supports a wider useful learning-rate range, not immunity to an excessive rate or elimination of task-specific trade-offs.

Table 24: Long-horizon summary for 1-shot CLIP (ViT-B/16). Final accuracy is measured at epoch 300. The best-to-final drop is computed separately within each seed before aggregation. For these two quantities, we average the 11 datasets within each seed and report mean ± population standard deviation over seeds 1, 2, and 3. The last column is the mean accuracy change from epoch 30 to epoch 300. All 528 runs are included. Higher final accuracy and smaller drops are desirable.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	LoRA	2.14 ± 0.27	54.80 ± 0.47	-0.06
	SO	7.04 ± 2.32	59.97 ± 2.11	-28.12
	SAFT	5.37 ± 0.19	41.26 ± 0.19	$+2.82$
	OLM-SGD	47.38 ± 0.43	3.45 ± 0.80	-0.18
10^{-3}	LoRA	65.98 ± 0.31	5.72 ± 0.24	-0.89
	SO	52.47 ± 0.16	20.02 ± 0.17	-15.58
	SAFT	53.51 ± 0.36	9.72 ± 0.12	-0.78
	OLM-SGD	69.23 ± 0.54	2.22 ± 0.07	-0.91
10^{-4}	LoRA	67.98 ± 0.36	3.47 ± 0.27	-1.76
	SO	68.51 ± 0.25	3.83 ± 0.24	-1.92
	SAFT	68.54 ± 0.50	2.92 ± 0.13	-1.43
	OLM-SGD	71.75 ± 0.26	0.95 ± 0.06	$+0.14$
10^{-5}	LoRA	69.77 ± 0.04	1.17 ± 0.06	$+4.16$
	SO	71.23 ± 0.11	0.82 ± 0.13	$+5.54$
	SAFT	69.88 ± 0.12	1.35 ± 0.08	$+0.26$
	OLM-SGD	71.75 ± 0.26	0.11 ± 0.00	$+3.85$

Training loss does not determine test accuracy. The loss curves separate support-set fitting from generalization. On SUN397 at 10^{-4}, all four methods reach 100% final training accuracy. Their test accuracies nevertheless differ: OLM-SGD reaches 70.10%, whereas the baselines range from 65.96% to 66.64%. Conversely, OLM-SGD need not attain the lowest training loss to obtain the best test result. On Aircraft at 10^{-5}, its final mean loss is 0.635 and its test accuracy is 30.50%. SAFT

attains a much lower loss of 0.036 but only 26.85% test accuracy. These observations are consistent with less harmful adaptation to the limited support set.

Numerical failures at the largest rate. LoRA produces nonfinite losses on ImageNet at 10^{-2} from epoch 102 for seed 1 and epoch 117 for seed 3. A further nonfinite segment occurs on SUN397 from epoch 296 for seed 3. These are numerical failures within completed 300-epoch logs, not unfinished trajectories. No other nonfinite loss entries occur in the 528 runs. All logged test accuracies remain finite. We retain the failed runs in the accuracy summaries rather than excluding them. Numerical stability and test-accuracy retention are distinct: finite losses alone do not make the 10^{-2} setting effective for OLM-SGD.

F.4 Best-to-Final Test-Accuracy Drop

Metric and aggregation. Tables 25–28 quantify the degradation visible in the accuracy curves. For each dataset and seed, the drop is the highest test accuracy over epochs 1–300 minus the accuracy at epoch 300. We then report mean and population standard deviation across the three seeds. The Average row first averages the 11 dataset-level drops within each seed. This differs from maximizing a seed-mean curve, as done in the epoch-transfer diagnostic. Maximization and seed averaging are not interchangeable. All differences are measured in percentage points.

Smallest mean degradation at every tested rate. OLM-SGD's mean drops are 3.45, 2.22, 0.95, and 0.11 points at 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}, respectively. Each is lower than the corresponding mean for LoRA, SO, and SAFT. At 10^{-4}, their drops are 3.47, 3.83, and 2.92 points. OLM-SGD stays within one point of its seed-specific best accuracy on nine datasets when drops are averaged over seeds. The corresponding counts are six for LoRA, five for SO, and six for SAFT.

Dataset-level retention and the ImageNet outlier. At 10^{-4}, OLM-SGD loses 0.72 points on Stanford Cars, compared with 5.00–6.39 for the baselines. Its SUN397 drop is 0.28 points, compared with 2.73–3.09. On Aircraft, the drop is 1.56 points rather than 4.57–5.05. ImageNet remains the main limitation. OLM-SGD loses 7.32 points there, despite improving substantially over the baseline drops of 16.69–22.62. Excluding ImageNet, its mean drop across the other ten datasets is 0.31 points. Thus, the aggregate 0.95-point result combines broadly stable trajectories with one substantially less stable target.

At 10^{-5}, OLM-SGD's dataset-level mean drops are all at most 0.43 points. Its overall drop is 0.11 points, compared with 1.17 for LoRA, 0.82 for SO, and 1.35 for SAFT. This retention is accompanied by 71.75% mean final accuracy, rather than a flat but ineffective trajectory. Nevertheless, as shown above, SO remains competitive at this rate and has higher final accuracy on several datasets.

A small drop is not sufficient evidence of successful adaptation. At 10^{-2}, SAFT's ImageNet drop is only 0.08 points, but its final mean accuracy is 0.51%. The metric is small because even the best recorded checkpoint performs poorly. Similarly, OLM-SGD's smaller mean drop at this rate does not offset its low absolute final accuracy. The metric also excludes the zero-shot checkpoint, so degradation before the first logged epoch is not captured. It must therefore be read jointly with final accuracy and the zero-shot reference. Taken together, the four analyses support a stronger accuracy–stability trade-off for OLM-SGD under the evaluated settings.

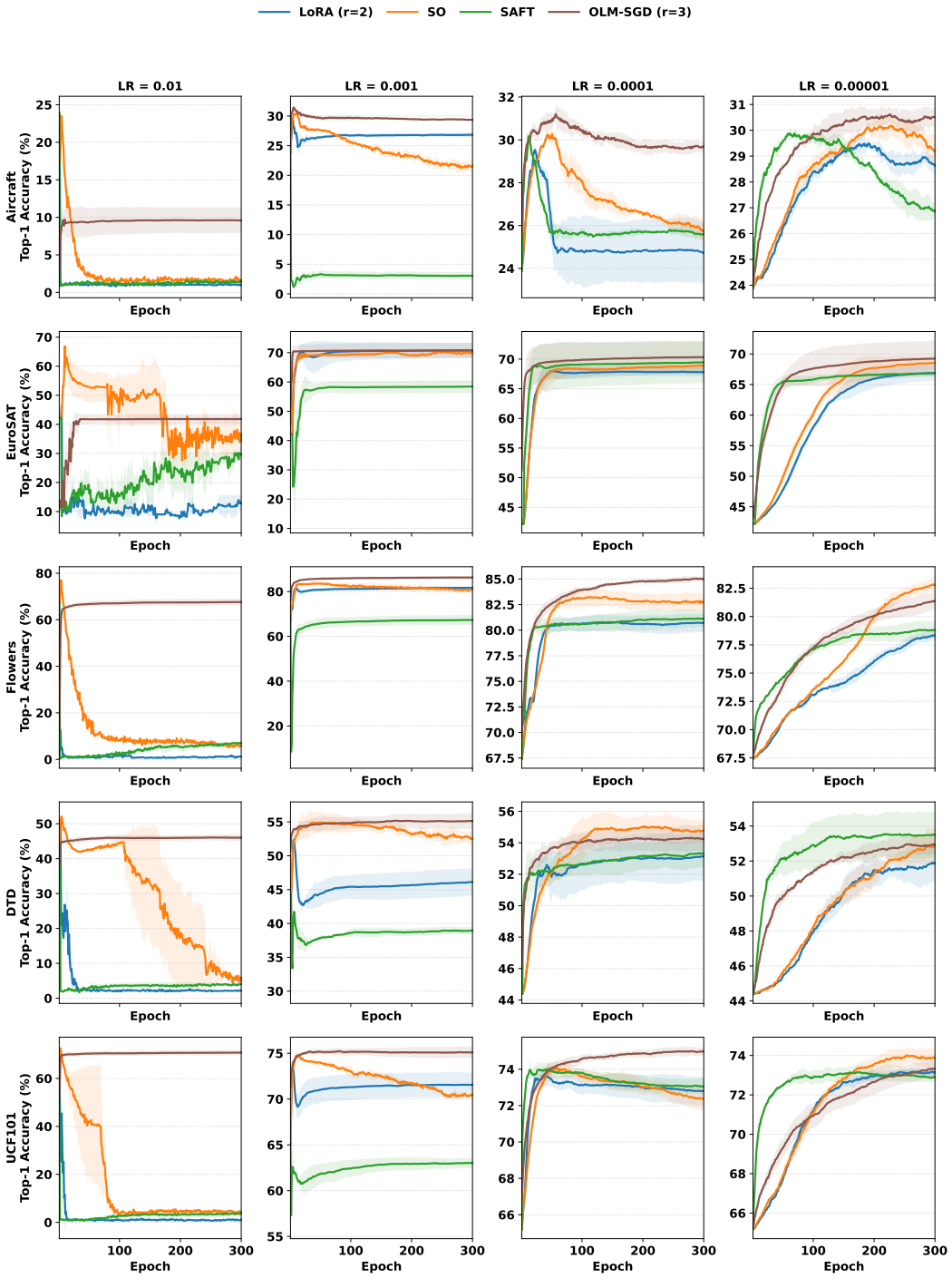

Figure 5: Long-horizon test accuracy for 1-shot CLIP (ViT-B/16) without data augmentation. Methods are LoRA (rank 2), SO and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one standard deviation over seeds 1, 2, and 3. All runs contain 300 epochs.

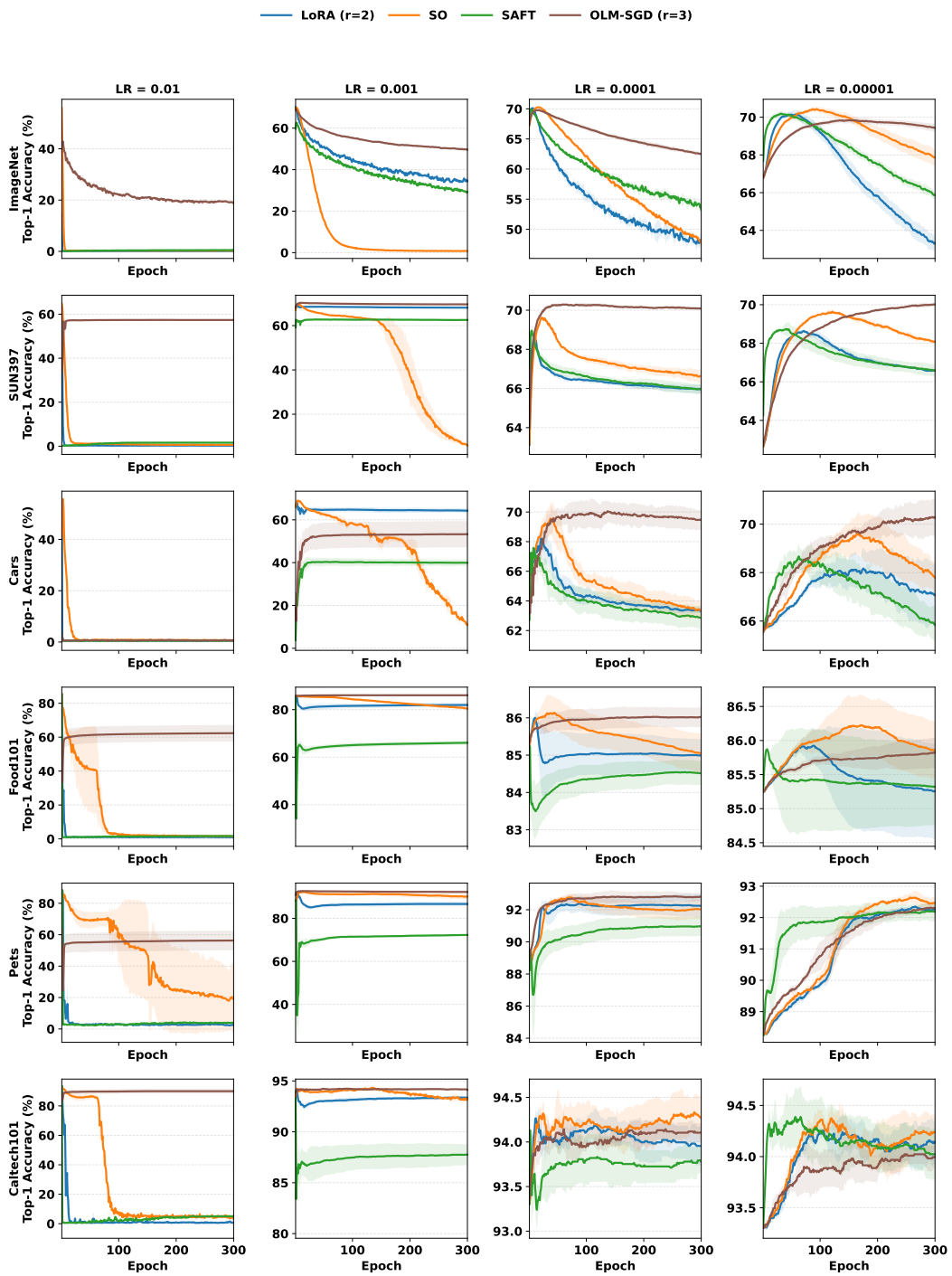

Figure 6: Long-horizon test accuracy for the remaining six datasets under the protocol of Figure 5. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. OLM-SGD retains higher final accuracy on most datasets at 10^{-4}, but ImageNet still exhibits substantial degradation. All runs contain 300 epochs.

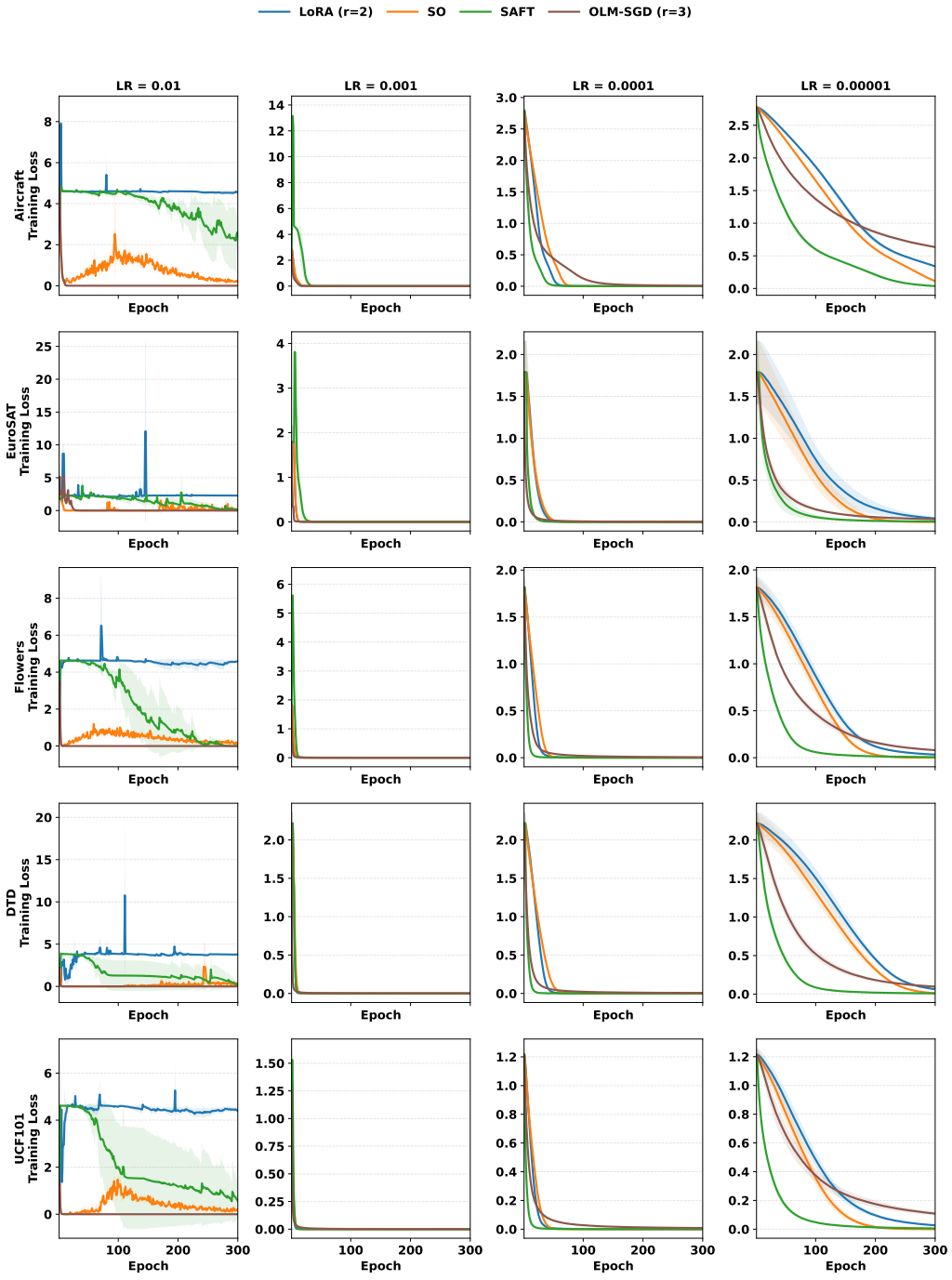

Figure 7: Training loss for the trajectories in Figure 5. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one standard deviation across three seeds. Small training loss can coexist with substantially different test accuracies.

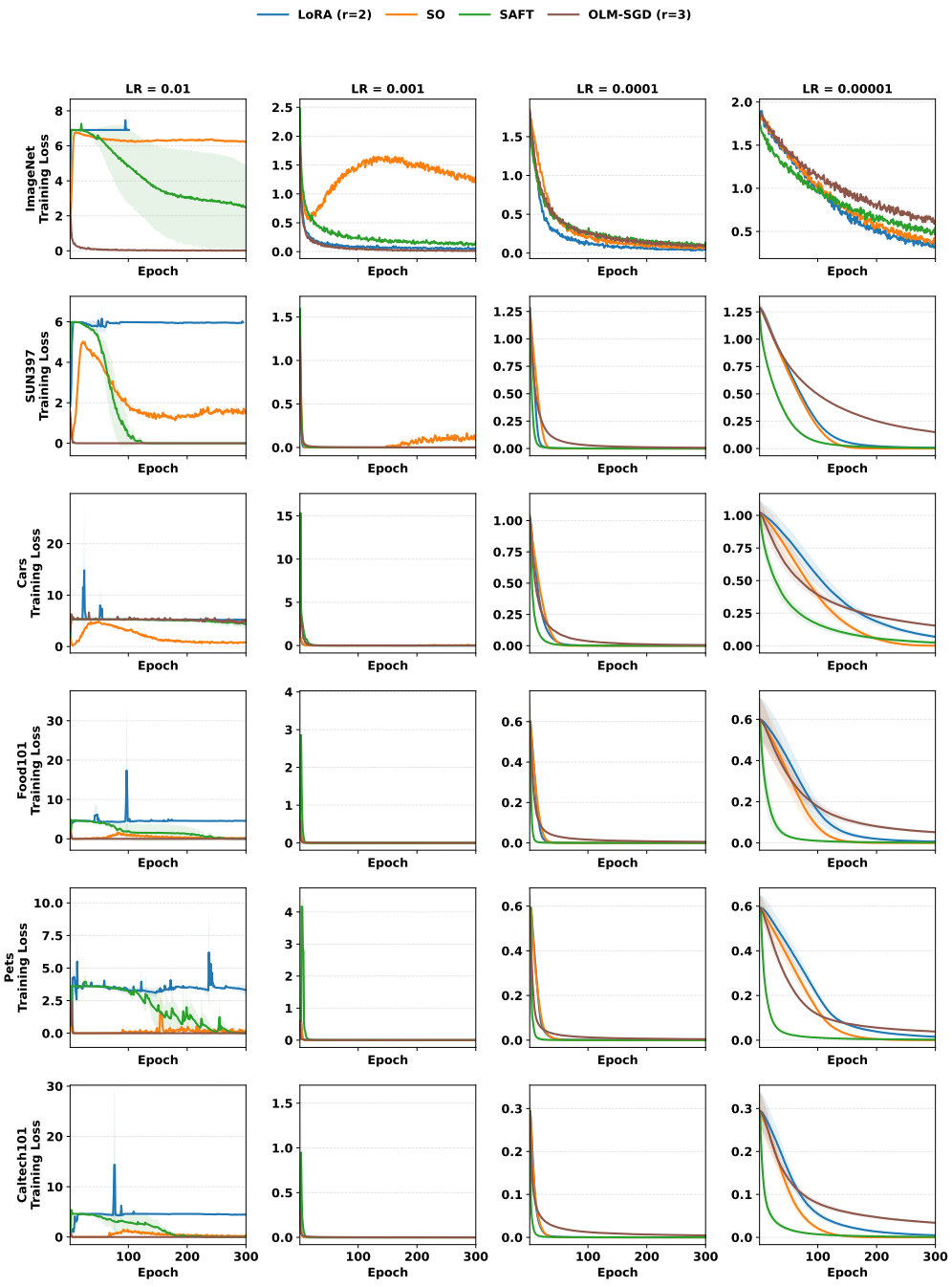

Figure 8: Training loss for the trajectories in Figure 6. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one standard deviation across three seeds. At 10^{-2}, LoRA has nonfinite losses on ImageNet from epochs 102 and 117 for seeds 1 and 3, respectively. Its SUN397 loss is also nonfinite from epoch 296 for seed 3. All three logs for each dataset contain 300 epochs. Missing loss segments indicate numerical failure rather than incomplete training.

Table 25: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) adaptation at learning rate 10^{-2}. Lower results are better. For each seed, the drop is the highest accuracy over epochs 1–300 minus the accuracy at epoch 300. Values are mean ± population standard deviation across seeds 1, 2, and 3. LoRA ($r = 2$), SO (99.9% sparsity) and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). The Average row first takes the unweighted mean over all 11 datasets within each seed, then reports the mean and standard deviation across seeds.

Dataset	LoRA	SO	SAFT	OLM-SGD
ImageNet	14.53 ± 4.09	55.65 ± 0.61	0.08 ± 0.01	24.69 ± 0.49
Caltech101	92.59 ± 0.11	89.29 ± 0.37	88.29 ± 1.11	0.07 ± 0.05
Oxford Pets	85.65 ± 0.38	69.27 ± 20.37	84.31 ± 0.86	9.29 ± 6.93
Stanford Cars	51.04 ± 3.81	55.06 ± 0.29	0.21 ± 0.17	2.83 ± 3.57
Flowers102	66.18 ± 0.29	71.13 ± 2.11	60.31 ± 1.05	0.04 ± 0.06
Food101	84.27 ± 0.05	83.64 ± 0.08	83.57 ± 0.17	0.00 ± 0.00
Aircraft	22.95 ± 0.07	22.63 ± 0.56	22.41 ± 0.25	0.57 ± 0.39
SUN397	38.94 ± 4.66	64.25 ± 0.48	0.06 ± 0.03	0.17 ± 0.17
DTD	49.19 ± 0.46	47.48 ± 1.82	40.37 ± 0.39	0.12 ± 0.08
EuroSAT	33.45 ± 6.10	33.13 ± 5.55	12.63 ± 1.59	0.17 ± 0.13
UCF101	64.02 ± 0.38	68.17 ± 0.55	61.60 ± 1.18	0.03 ± 0.04
Average	54.80 ± 0.47	59.97 ± 2.11	41.26 ± 0.19	3.45 ± 0.80

Table 26: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) adaptation at learning rate 10^{-3}. Lower results are better. For each seed, the drop is the highest accuracy over epochs 1–300 minus the accuracy at epoch 300. Values are mean ± population standard deviation across seeds 1, 2, and 3. LoRA ($r = 2$), SO (99.9% sparsity) and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). The Average row first takes the unweighted mean over all 11 datasets within each seed, then reports the mean and standard deviation across seeds.

Dataset	LoRA	SO	SAFT	OLM-SGD
ImageNet	35.79 ± 0.31	69.09 ± 0.20	33.42 ± 0.49	19.37 ± 0.63
Caltech101	0.93 ± 0.34	1.15 ± 0.08	5.54 ± 1.01	0.20 ± 0.09
Oxford Pets	5.07 ± 0.62	2.22 ± 0.56	15.99 ± 0.34	0.45 ± 0.15
Stanford Cars	3.81 ± 1.00	58.15 ± 0.86	0.89 ± 0.66	0.03 ± 0.02
Flowers102	0.47 ± 0.64	2.80 ± 0.52	1.15 ± 0.93	0.05 ± 0.02
Food101	4.04 ± 0.98	5.65 ± 0.63	19.20 ± 0.81	0.05 ± 0.02
Aircraft	2.94 ± 0.68	9.17 ± 0.94	20.82 ± 0.64	2.20 ± 0.39
SUN397	1.12 ± 0.49	63.92 ± 0.60	0.55 ± 0.28	0.74 ± 0.18
DTD	6.40 ± 0.70	2.78 ± 0.89	5.42 ± 0.58	0.14 ± 0.20
EuroSAT	0.37 ± 0.33	0.84 ± 0.80	1.46 ± 1.07	0.84 ± 0.18
UCF101	1.95 ± 1.20	4.48 ± 0.32	2.43 ± 0.66	0.29 ± 0.23
Average	5.72 ± 0.24	20.02 ± 0.17	9.72 ± 0.12	2.22 ± 0.07

Table 27: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) adaptation at learning rate 10^{-4}. Lower results are better. For each seed, the drop is the highest accuracy over epochs 1–300 minus the accuracy at epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. LoRA ($r = 2$), SO (99.9% sparsity) and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). The Average row first takes the unweighted mean over all 11 datasets within each seed, then reports the mean and standard deviation across seeds.

Dataset	LoRA	SO	SAFT	OLM-SGD
ImageNet	22.01 ± 0.52	22.62 ± 0.20	16.69 ± 0.70	7.32 ± 0.34
Caltech101	0.42 ± 0.28	0.22 ± 0.13	0.41 ± 0.18	0.16 ± 0.10
Oxford Pets	0.31 ± 0.25	0.82 ± 0.50	0.03 ± 0.02	0.07 ± 0.05
Stanford Cars	5.00 ± 0.74	6.39 ± 1.18	5.07 ± 0.75	0.72 ± 0.13
Flowers102	0.16 ± 0.14	0.60 ± 0.10	0.05 ± 0.08	0.04 ± 0.00
Food101	1.05 ± 0.34	1.26 ± 0.29	0.80 ± 0.26	0.02 ± 0.02
Aircraft	5.05 ± 1.55	4.57 ± 0.25	4.69 ± 0.10	1.56 ± 0.22
SUN397	2.73 ± 0.39	3.02 ± 0.37	3.09 ± 0.17	0.28 ± 0.04
DTD	0.14 ± 0.11	0.47 ± 0.29	0.06 ± 0.00	0.20 ± 0.17
EuroSAT	0.40 ± 0.31	0.40 ± 0.52	0.21 ± 0.14	0.01 ± 0.02
UCF101	0.91 ± 0.29	1.77 ± 0.27	1.04 ± 0.22	0.04 ± 0.04
Average	3.47 ± 0.27	3.83 ± 0.24	2.92 ± 0.13	0.95 ± 0.06

Table 28: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) adaptation at learning rate 10^{-5}. Lower results are better. For each seed, the drop is the highest accuracy over epochs 1–300 minus the accuracy at epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. LoRA ($r = 2$), SO (99.9% sparsity) and SAFT (99.9% sparsity), and OLM-SGD ($r = 3$). The Average row first takes the unweighted mean over all 11 datasets within each seed, then reports the mean and standard deviation across seeds.

Dataset	LoRA	SO	SAFT	OLM-SGD
ImageNet	6.87 ± 0.23	2.57 ± 0.29	4.34 ± 0.14	0.43 ± 0.06
Caltech101	0.19 ± 0.12	0.24 ± 0.11	0.50 ± 0.27	0.07 ± 0.04
Oxford Pets	0.11 ± 0.07	0.25 ± 0.14	0.05 ± 0.08	0.02 ± 0.03
Stanford Cars	1.33 ± 0.56	1.95 ± 0.58	2.97 ± 0.27	0.19 ± 0.16
Flowers102	0.03 ± 0.02	0.08 ± 0.03	0.04 ± 0.03	0.01 ± 0.02
Food101	0.78 ± 0.52	0.46 ± 0.14	0.68 ± 0.47	0.06 ± 0.09
Aircraft	1.02 ± 0.37	1.22 ± 0.35	3.29 ± 0.39	0.29 ± 0.21
SUN397	2.12 ± 0.09	1.57 ± 0.03	2.27 ± 0.07	0.02 ± 0.00
DTD	0.35 ± 0.17	0.20 ± 0.15	0.16 ± 0.18	0.08 ± 0.06
EuroSAT	0.01 ± 0.01	0.23 ± 0.30	0.19 ± 0.13	0.00 ± 0.01
UCF101	0.11 ± 0.10	0.19 ± 0.06	0.35 ± 0.19	0.01 ± 0.01
Average	1.17 ± 0.06	0.82 ± 0.13	1.35 ± 0.08	0.11 ± 0.00

G Comparison against Robust Fine-Tuning Methods

Comparison setting. We compare OLM-SGD with three methods that preserve pretrained knowledge during adaptation. VRF (Zhu et al., 2024) combines zero-shot and fine-tuned predictions through input-dependent weights. WiSE (Wortsman et al., 2022b) interpolates pretrained and fine-tuned parameters. CLIPood (Shu et al., 2023) combines margin metric softmax with Beta moving average. We evaluate each method with Adam and SGD while keeping its method-specific hyperparameters fixed. This separates the effect of the base optimizer from the choice of robust fine-tuning method. Unlike these methods, OLM-SGD modifies the gradient-merging rule without adding prediction ensembling, checkpoint interpolation, or a modified classification loss.

Evaluation protocol. Tables 30, 31, and 32 report results for CLIP with ViT-B/32, ViT-B/16, and ViT-L/14. We evaluate $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class on all 11 datasets using seeds 1, 2, and 3. Entries report mean top-1 test accuracy and population standard deviation across seeds. The *Average* column first averages the 11 dataset accuracies within each seed. It then reports the mean and standard deviation of these three dataset averages. Thus, each dataset receives equal weight regardless of its test-set size. The optimization settings are given in Appendix D. All methods adapt the vision and text Transformer residual blocks at learning rate 10^{-4}. The robust baselines use 30 epochs and OLM-SGD uses 70. These results compare the stated adaptation schedules.

Overall accuracy. OLM-SGD attains the highest dataset-average accuracy in 14 of the 15 backbone–shot settings. VRF with SGD is the strongest robust baseline in all 15 settings. Table 29 summarizes this comparison. On ViT-B/32, OLM-SGD improves from 67.2% at 1 shot to 78.5% at 16 shots. Its gains over VRF-SGD are 0.1, 0.5, 0.7, 0.8, and 0.8 percentage points at the five shot levels. On ViT-B/16, OLM-SGD reaches 72.0%, 75.1%, 77.8%, 80.3%, and 82.6%. The corresponding gains are 0.4, 0.5, 0.9, 1.1, and 0.9 points. On ViT-L/14, OLM-SGD reaches 77.4%, 80.1%, 82.5%, 84.7%, and 86.8%. It exceeds VRF-SGD at 2–16 shots but trails it by 0.3 points at 1 shot. The average gains over VRF-SGD across shot levels are approximately 0.58, 0.76, and 0.22 points for the 3 backbones. The advantage is consistent in the aggregate but smaller on ViT-L/14.

Table 29: Dataset-average top-1 accuracy (%) for OLM-SGD and the strongest robust baseline, VRF-SGD. Each accuracy is the corresponding *Average* entry in Tables 30–32. The Δ rows report OLM-SGD minus VRF-SGD in percentage points, computed from the displayed means. Bold marks the higher accuracy within each backbone–shot pair.

Backbone	Method	1 shot	2 shots	4 shots	8 shots	16 shots
ViT-B/32	VRF (SGD)	67.1	69.9	72.4	75.0	77.7
	OLM-SGD	**67.2**	**70.4**	**73.1**	**75.8**	**78.5**
	Δ	$+0.1$	$+0.5$	$+0.7$	$+0.8$	$+0.8$
ViT-B/16	VRF (SGD)	71.6	74.6	76.9	79.2	81.7
	OLM-SGD	**72.0**	**75.1**	**77.8**	**80.3**	**82.6**
	Δ	$+0.4$	$+0.5$	$+0.9$	$+1.1$	$+0.9$
ViT-L/14	VRF (SGD)	**77.7**	80.0	82.0	84.4	86.3
	OLM-SGD	77.4	**80.1**	**82.5**	**84.7**	**86.8**
	Δ	-0.3	$+0.1$	$+0.5$	$+0.3$	$+0.5$

The base optimizer remains decisive. Replacing Adam with SGD improves the dataset-average accuracy of each robust method in every backbone–shot setting. Averaged over the 15 settings, the gains are 38.7 points for VRF, 40.8 for WiSE, and 25.4 for CLIPood. For example, at 1 shot with ViT-B/16, VRF improves from 30.3% to 71.6%. WiSE improves from 22.4% to 69.9%, and CLIPood improves from 37.3% to 69.7%. All three Adam variants remain below the corresponding zero-shot dataset average at every shot level and backbone. Thus, prediction ensembling, parameter interpolation, and temporal averaging do not eliminate sensitivity to the underlying optimizer.

Comparison with the SGD variants. The gains of OLM-SGD are not explained only by the low Adam results. Across shot levels, its mean advantages over WiSE-SGD are approximately 3.8, 3.9, and 3.4 points on ViT-B/32, ViT-B/16, and ViT-L/14. The corresponding advantages over CLIPood-SGD are 3.9, 3.5, and 3.2 points. VRF-SGD is substantially more competitive, which narrows the remaining gap to OLM-SGD. Nevertheless, OLM-SGD leads it in all 12 settings with at least two

shots per class. At 16 shots, OLM-SGD attains dataset averages of 78.5%, 82.6%, and 86.8%, compared with 77.7%, 81.7%, and 86.3% for VRF-SGD. Its gains over zero-shot CLIP at this shot level are 18.3, 18.5, and 15.5 points. The benefit extends beyond the most label-scarce setting.

Dataset-level improvements. Across the 165 dataset–backbone–shot settings, OLM-SGD is strictly best against the six robust variants in 99 settings and tied for best in 12. Its gains are distributed across several datasets rather than concentrated in a single benchmark. OLM-SGD leads on Oxford Flowers in all 15 settings. On DTD, it leads in 14 settings and ties in the remaining one. It also leads in 13 settings on both SUN397 and EuroSAT. On Stanford Cars, it leads in 13 settings and ties in one. FGVC-Aircraft and UCF101 each contribute 12 strict wins.

Several gains become more pronounced as the support set grows. With ViT-B/16 on Aircraft, the margin over VRF-SGD is 1.3 points at 1 shot and 4.2 points at 16 shots. At 16 shots, OLM-SGD reaches 52.9% on Aircraft and 83.4% on Cars, compared with 48.7% and 81.4% for VRF-SGD. On ViT-B/32, its 4-shot results on Flowers and DTD are 89.6% and 61.1%. The corresponding VRF-SGD results are 87.1% and 58.4%. The gains also persist on the larger backbone. For example, 16-shot ViT-L/14 accuracy on EuroSAT improves from 91.8% with VRF-SGD to 93.7% with OLM-SGD. These comparisons show that the aggregate advantage is supported by improvements on multiple datasets and shot levels.

Where robust fine-tuning remains stronger. ImageNet is the clearest exception. OLM-SGD trails the strongest robust baseline in all 15 ImageNet settings. At 16 shots, it obtains 64.4% on ViT-B/32, whereas WiSE-SGD reaches 68.6%. On ViT-B/16, it obtains 70.8%, compared with 73.4% for both WiSE-SGD and CLIPood-SGD. On ViT-L/14, it obtains 79.2%, compared with 79.8% for VRF-SGD. The 11-dataset average therefore does not imply uniform superiority. On Caltech101, OLM-SGD leads in only one setting and ties in two. Food101 and Oxford Pets also show smaller differences and several ties. In particular, OLM-SGD matches the highest 16-shot Food101 accuracy on all three backbones. These results indicate that the robust baselines retain useful dataset-specific strengths even when OLM-SGD has a higher overall average.

Table 30: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. OLM-SGD is compared with robust full-finetuning methods VRF, WiSE, and CLIPood, each trained with Adam and SGD. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	VRF (Adam)	11.0 ± 1.7	23.6 ± 2.0	12.4 ± 8.1	25.9 ± 11.9	40.3 ± 28.2	29.5 ± 35.1	12.6 ± 2.4	50.3 ± 14.6	56.3 ± 2.0	17.2 ± 2.0	25.4 ± 1.9	27.7 ± 1.3
	VRF (SGD)	62.1 ± 0.3	67.2 ± 0.2	22.4 ± 0.4	62.3 ± 5.4	62.0 ± 0.4	80.5 ± 0.2	89.2 ± 0.6	77.9 ± 0.9	93.7 ± 0.6	50.1 ± 0.0	70.2 ± 0.5	67.1 ± 0.4
	WiSE (Adam)	28.8 ± 3.7	45.8 ± 1.6	1.0 ± 0.0	9.9 ± 0.9	0.5 ± 0.0	3.2 ± 3.1	12.1 ± 4.1	3.4 ± 2.4	75.8 ± 2.4	26.5 ± 1.7	49.4 ± 5.3	23.3 ± 0.4
	WiSE (SGD)	64.2 ± 0.3	66.6 ± 0.2	21.1 ± 0.4	54.1 ± 3.4	63.0 ± 0.3	80.6 ± 0.0	87.7 ± 0.2	71.9 ± 0.2	93.0 ± 0.4	49.1 ± 0.5	67.3 ± 0.1	65.3 ± 0.4
	CLIPood (Adam)	39.8 ± 0.5	48.1 ± 0.7	2.2 ± 0.5	35.2 ± 4.8	0.6 ± 0.0	12.3 ± 4.4	58.4 ± 1.7	45.2 ± 6.9	84.2 ± 1.0	31.4 ± 2.1	51.5 ± 3.6	37.2 ± 0.8
	CLIPood (SGD)	63.6 ± 0.0	64.9 ± 0.1	21.3 ± 0.7	61.0 ± 0.8	60.6 ± 0.2	80.0 ± 0.2	86.7 ± 0.4	68.5 ± 0.7	92.6 ± 0.2	46.6 ± 0.2	65.1 ± 0.2	64.6 ± 0.2
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.6	50.7 ± 1.9	70.1 ± 0.4	67.2 ± 0.5
2	VRF (Adam)	14.5 ± 1.0	22.3 ± 0.8	18.0 ± 0.1	28.0 ± 12.9	60.2 ± 0.1	29.3 ± 35.1	30.2 ± 29.6	35.8 ± 25.8	67.5 ± 2.6	19.7 ± 5.0	25.1 ± 3.3	31.9 ± 6.3
	VRF (SGD)	62.5 ± 0.2	68.4 ± 0.3	23.9 ± 0.1	75.0 ± 2.3	63.8 ± 0.1	80.8 ± 0.1	89.8 ± 0.6	82.6 ± 0.7	94.5 ± 0.1	54.7 ± 1.6	72.6 ± 0.4	69.9 ± 0.1
	WiSE (Adam)	32.8 ± 0.5	42.7 ± 2.8	1.0 ± 0.0	15.5 ± 4.8	0.4 ± 0.0	2.6 ± 2.2	14.2 ± 14.3	1.3 ± 0.9	81.8 ± 3.3	27.4 ± 9.6	42.1 ± 5.9	23.8 ± 1.7
	WiSE (SGD)	65.1 ± 0.1	67.5 ± 0.1	22.8 ± 0.4	62.0 ± 1.7	64.2 ± 0.2	81.0 ± 0.1	88.1 ± 0.4	74.8 ± 0.5	93.8 ± 0.2	52.9 ± 1.0	69.3 ± 0.2	67.4 ± 0.0
	CLIPood (Adam)	40.2 ± 0.0	53.5 ± 0.4	2.9 ± 2.2	24.4 ± 19.2	0.7 ± 0.1	6.2 ± 2.9	51.9 ± 16.2	46.8 ± 5.7	85.1 ± 1.6	36.6 ± 2.6	56.4 ± 1.7	36.8 ± 2.9
	CLIPood (SGD)	64.2 ± 0.2	65.6 ± 0.3	22.6 ± 0.4	66.1 ± 3.1	61.7 ± 0.5	79.9 ± 0.1	87.6 ± 0.2	72.6 ± 0.5	93.4 ± 0.4	49.5 ± 0.4	66.9 ± 0.5	66.4 ± 0.2
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.9 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.4 ± 0.9	70.4 ± 0.1
4	VRF (Adam)	19.1 ± 0.7	24.6 ± 0.8	15.6 ± 2.0	35.5 ± 2.4	42.1 ± 20.9	46.5 ± 32.1	32.4 ± 10.7	19.9 ± 12.9	66.1 ± 6.3	25.6 ± 3.1	27.0 ± 6.7	32.2 ± 0.1
	VRF (SGD)	63.7 ± 0.2	69.8 ± 0.3	27.5 ± 0.3	80.8 ± 1.4	66.6 ± 0.4	81.0 ± 0.2	90.8 ± 0.3	87.1 ± 0.6	94.9 ± 0.3	58.4 ± 0.6	75.5 ± 0.7	72.4 ± 0.2
	WiSE (Adam)	36.3 ± 1.0	49.5 ± 0.5	1.3 ± 0.3	22.6 ± 4.2	1.1 ± 0.5	2.4 ± 1.4	23.6 ± 13.8	15.6 ± 11.9	83.1 ± 5.0	33.9 ± 4.3	35.7 ± 12.6	27.7 ± 1.9
	WiSE (SGD)	66.1 ± 0.1	68.7 ± 0.1	24.6 ± 0.3	66.3 ± 0.6	65.7 ± 0.3	81.3 ± 0.2	88.6 ± 0.2	77.6 ± 0.1	94.3 ± 0.4	55.0 ± 0.6	70.6 ± 0.1	69.0 ± 0.1
	CLIPood (Adam)	42.3 ± 0.5	55.7 ± 1.2	3.2 ± 1.4	28.0 ± 12.5	4.8 ± 2.8	22.2 ± 6.3	64.7 ± 2.1	65.8 ± 2.2	90.4 ± 1.5	40.7 ± 3.4	63.6 ± 1.4	43.8 ± 1.4
	CLIPood (SGD)	65.1 ± 0.1	66.8 ± 0.0	24.8 ± 0.7	69.0 ± 1.9	63.2 ± 0.0	80.4 ± 0.3	88.4 ± 0.3	76.4 ± 0.4	93.7 ± 0.1	51.6 ± 0.3	68.6 ± 0.5	68.0 ± 0.2
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.1 ± 0.6	81.4 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1 ± 0.2
8	VRF (Adam)	24.5 ± 0.2	29.5 ± 0.5	9.4 ± 4.2	31.5 ± 4.8	30.0 ± 23.0	53.1 ± 32.4	48.2 ± 24.0	37.1 ± 12.9	73.5 ± 5.6	29.0 ± 5.5	38.1 ± 2.6	36.7 ± 4.1
	VRF (SGD)	64.9 ± 0.2	71.1 ± 0.2	32.5 ± 0.9	88.4 ± 0.7	70.5 ± 0.3	81.2 ± 0.1	91.2 ± 0.3	90.4 ± 0.5	95.2 ± 0.4	62.5 ± 0.5	77.7 ± 0.1	75.0 ± 0.2
	WiSE (Adam)	39.6 ± 0.3	53.2 ± 0.4	1.5 ± 0.2	19.6 ± 6.0	1.4 ± 0.6	4.3 ± 2.7	7.6 ± 4.1	34.0 ± 15.4	88.4 ± 2.6	35.0 ± 7.9	58.9 ± 2.4	31.2 ± 0.7
	WiSE (SGD)	67.1 ± 0.2	70.0 ± 0.0	27.0 ± 0.2	74.4 ± 1.2	68.4 ± 0.2	81.6 ± 0.0	89.4 ± 0.3	80.2 ± 0.1	94.6 ± 0.2	57.2 ± 0.3	72.3 ± 0.4	71.1 ± 0.1
	CLIPood (Adam)	44.8 ± 0.2	56.8 ± 0.0	9.9 ± 3.0	39.8 ± 4.6	2.7 ± 1.8	29.5 ± 4.1	68.9 ± 6.0	79.9 ± 2.3	92.6 ± 0.8	49.4 ± 0.5	70.6 ± 1.0	49.5 ± 0.9
	CLIPood (SGD)	66.2 ± 0.1	68.1 ± 0.2	28.2 ± 0.2	85.0 ± 0.7	65.1 ± 0.2	80.4 ± 0.2	90.0 ± 0.1	83.2 ± 0.2	94.5 ± 0.1	56.2 ± 0.1	71.1 ± 0.2	71.6 ± 0.1
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.6 ± 0.1	91.2 ± 0.5	91.9 ± 0.1	95.2 ± 0.2	64.6 ± 0.6	78.5 ± 0.4	75.8 ± 0.1
16	VRF (Adam)	29.5 ± 0.6	32.6 ± 1.1	4.2 ± 0.2	47.4 ± 16.2	19.3 ± 20.6	9.5 ± 2.9	23.8 ± 1.8	52.6 ± 15.1	79.7 ± 2.8	38.8 ± 2.2	33.2 ± 16.0	33.7 ± 3.5
	VRF (SGD)	67.4 ± 0.2	72.5 ± 0.1	39.5 ± 0.1	91.9 ± 0.4	74.7 ± 0.3	81.5 ± 0.0	91.1 ± 0.3	93.9 ± 0.2	95.5 ± 0.0	67.1 ± 0.7	79.8 ± 0.1	77.7 ± 0.1
	WiSE (Adam)	38.8 ± 1.7	54.6 ± 0.3	2.0 ± 0.1	33.8 ± 9.0	3.9 ± 2.2	11.2 ± 4.4	26.0 ± 2.1	49.4 ± 21.4	91.2 ± 1.0	48.4 ± 4.6	43.7 ± 27.0	36.6 ± 2.8
	WiSE (SGD)	68.6 ± 0.1	71.3 ± 0.1	30.2 ± 0.4	79.3 ± 1.0	71.0 ± 0.3	81.9 ± 0.0	89.7 ± 0.0	83.1 ± 0.4	95.0 ± 0.1	60.4 ± 0.2	74.3 ± 0.3	73.2 ± 0.1
	CLIPood (Adam)	47.8 ± 0.2	58.9 ± 0.3	15.9 ± 2.7	50.7 ± 6.2	5.6 ± 0.8	46.9 ± 8.6	70.7 ± 3.3	89.9 ± 0.8	94.1 ± 0.2	55.3 ± 0.9	73.8 ± 0.4	55.4 ± 1.2
	CLIPood (SGD)	67.7 ± 0.1	70.0 ± 0.0	32.6 ± 0.3	91.3 ± 0.4	68.8 ± 0.2	80.8 ± 0.1	91.3 ± 0.3	90.0 ± 0.1	95.4 ± 0.2	60.8 ± 0.1	74.5 ± 0.3	74.8 ± 0.1
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5 ± 0.1

Table 31: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. OLM-SGD is compared with robust full-finetuning methods VRF, WiSE, and CLIPood, each trained with Adam and SGD. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	VRF (Adam)	13.8 ± 0.9	24.0 ± 3.3	15.0 ± 9.9	27.1 ± 10.5	22.2 ± 30.7	50.0 ± 35.8	58.6 ± 39.8	23.0 ± 31.5	56.7 ± 4.2	21.0 ± 2.8	22.3 ± 6.0	30.3 ± 5.4
	VRF (SGD)	68.2 ± 0.2	69.9 ± 0.2	29.4 ± 0.3	68.2 ± 0.7	69.6 ± 0.9	86.1 ± 0.5	92.6 ± 0.3	82.5 ± 0.2	94.3 ± 0.1	53.2 ± 0.7	73.9 ± 0.2	71.6 ± 0.1
	WiSE (Adam)	34.2 ± 2.5	46.7 ± 2.4	1.1 ± 0.3	9.9 ± 1.7	0.5 ± 0.2	1.1 ± 0.1	5.7 ± 4.2	0.9 ± 0.4	77.4 ± 5.2	31.4 ± 2.5	37.1 ± 14.8	22.4 ± 0.8
	WiSE (SGD)	69.6 ± 0.2	68.4 ± 0.1	28.3 ± 0.3	63.9 ± 2.3	69.6 ± 0.3	86.1 ± 0.2	90.6 ± 0.3	76.4 ± 0.4	94.1 ± 0.1	51.1 ± 0.6	71.3 ± 0.5	69.9 ± 0.3
	CLIPood (Adam)	46.6 ± 0.3	52.1 ± 0.8	1.1 ± 0.1	35.3 ± 12.4	0.5 ± 0.0	19.1 ± 22.9	64.5 ± 5.8	14.4 ± 4.4	86.2 ± 0.6	35.1 ± 1.7	55.8 ± 3.3	37.3 ± 1.5
	CLIPood (SGD)	68.8 ± 0.1	67.7 ± 0.0	27.4 ± 0.4	70.8 ± 1.3	67.2 ± 0.4	85.9 ± 0.2	90.6 ± 0.3	73.8 ± 0.8	94.7 ± 0.0	49.3 ± 0.5	70.1 ± 0.4	69.7 ± 0.1
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.3	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0 ± 0.3
2	VRF (Adam)	29.5 ± 14.7	25.6 ± 3.5	8.7 ± 10.8	41.3 ± 0.3	22.2 ± 30.6	29.1 ± 39.7	31.2 ± 40.2	42.4 ± 20.9	67.6 ± 5.5	27.2 ± 8.1	31.0 ± 26.4	32.3 ± 11.2
	VRF (SGD)	68.5 ± 0.2	71.6 ± 0.1	32.5 ± 1.2	77.8 ± 1.2	71.8 ± 0.4	86.2 ± 0.1	92.8 ± 0.4	88.6 ± 0.5	94.6 ± 0.3	59.6 ± 1.1	77.0 ± 0.2	74.6 ± 0.2
	WiSE (Adam)	27.1 ± 18.0	49.4 ± 4.5	1.0 ± 0.0	12.5 ± 1.9	0.6 ± 0.0	1.1 ± 0.1	2.7 ± 0.0	4.9 ± 4.7	85.0 ± 1.2	17.4 ± 8.8	3.9 ± 4.3	18.7 ± 2.3
	WiSE (SGD)	70.1 ± 0.1	70.0 ± 0.1	30.9 ± 0.8	68.9 ± 1.8	71.0 ± 0.3	86.5 ± 0.2	91.2 ± 0.3	79.9 ± 0.2	94.4 ± 0.2	54.9 ± 1.1	73.1 ± 0.3	71.9 ± 0.3
	CLIPood (Adam)	47.7 ± 0.2	57.5 ± 0.5	1.1 ± 0.2	28.3 ± 4.4	0.8 ± 0.3	1.5 ± 0.5	31.0 ± 25.2	18.3 ± 6.4	87.1 ± 0.2	38.6 ± 4.2	59.8 ± 1.1	33.8 ± 1.9
	CLIPood (SGD)	69.5 ± 0.1	68.7 ± 0.2	29.5 ± 1.0	76.2 ± 1.3	68.3 ± 0.2	86.1 ± 0.1	90.9 ± 0.1	78.5 ± 0.4	94.9 ± 0.1	52.6 ± 0.6	71.7 ± 0.1	71.5 ± 0.2
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.2 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1 ± 0.2
4	VRF (Adam)	22.9 ± 0.8	30.5 ± 1.0	22.7 ± 1.5	42.1 ± 0.8	64.4 ± 1.6	32.1 ± 37.8	40.7 ± 30.5	31.3 ± 5.2	77.4 ± 1.0	20.5 ± 0.9	32.1 ± 10.0	37.9 ± 7.3
	VRF (SGD)	69.5 ± 0.1	72.6 ± 0.2	36.4 ± 0.6	82.6 ± 1.3	74.9 ± 0.2	86.6 ± 0.2	93.6 ± 0.3	91.8 ± 0.5	95.2 ± 0.2	63.3 ± 0.6	79.3 ± 0.4	76.9 ± 0.1
	WiSE (Adam)	42.0 ± 1.5	56.2 ± 0.9	1.0 ± 0.1	14.4 ± 3.7	0.8 ± 0.3	4.2 ± 4.1	18.7 ± 21.3	10.2 ± 2.6	90.1 ± 0.4	22.8 ± 4.0	42.9 ± 15.8	27.6 ± 2.0
	WiSE (SGD)	71.1 ± 0.1	71.0 ± 0.1	33.4 ± 0.5	73.2 ± 1.5	73.1 ± 0.1	86.7 ± 0.0	91.9 ± 0.1	81.6 ± 0.3	94.7 ± 0.1	56.9 ± 0.3	75.0 ± 0.4	73.5 ± 0.1
	CLIPood (Adam)	50.3 ± 0.4	60.9 ± 0.4	2.1 ± 0.5	21.8 ± 3.2	1.5 ± 0.4	9.6 ± 5.9	65.4 ± 8.8	56.3 ± 8.5	92.0 ± 0.4	46.3 ± 2.8	51.0 ± 21.4	41.6 ± 2.6
	CLIPood (SGD)	70.6 ± 0.1	69.9 ± 0.1	32.3 ± 0.3	80.9 ± 0.4	70.8 ± 0.2	86.3 ± 0.1	92.0 ± 0.2	82.6 ± 0.6	95.7 ± 0.1	54.4 ± 0.1	73.0 ± 0.3	73.5 ± 0.1
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.0	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8 ± 0.1
8	VRF (Adam)	29.7 ± 0.3	35.2 ± 1.6	20.6 ± 2.1	45.0 ± 5.6	63.7 ± 2.6	57.3 ± 39.5	47.7 ± 29.4	34.8 ± 9.9	73.1 ± 12.2	34.9 ± 1.9	38.1 ± 14.2	43.6 ± 3.5
	VRF (SGD)	70.7 ± 0.4	73.8 ± 0.3	42.3 ± 0.5	88.3 ± 0.6	77.9 ± 0.4	86.9 ± 0.2	93.6 ± 0.4	94.1 ± 0.3	95.5 ± 0.1	66.4 ± 0.7	81.4 ± 0.4	79.2 ± 0.3
	WiSE (Adam)	46.4 ± 0.4	58.9 ± 1.6	1.8 ± 0.2	11.5 ± 0.5	0.9 ± 0.4	1.1 ± 0.1	6.1 ± 2.6	25.4 ± 10.3	78.9 ± 18.6	40.7 ± 7.2	47.9 ± 22.4	29.1 ± 3.0
	WiSE (SGD)	72.3 ± 0.1	72.3 ± 0.2	35.9 ± 0.3	78.7 ± 0.7	75.2 ± 0.3	87.1 ± 0.1	92.7 ± 0.5	83.5 ± 0.1	94.9 ± 0.2	59.8 ± 0.4	76.7 ± 0.2	75.4 ± 0.2
	CLIPood (Adam)	53.1 ± 0.1	63.0 ± 0.3	2.5 ± 0.5	40.6 ± 5.6	12.8 ± 15.7	34.5 ± 10.9	52.6 ± 35.1	72.4 ± 6.9	94.1 ± 0.2	54.4 ± 1.2	75.5 ± 2.3	50.5 ± 3.4
	CLIPood (SGD)	71.8 ± 0.1	71.2 ± 0.1	35.8 ± 0.4	87.5 ± 0.8	73.6 ± 0.1	86.6 ± 0.1	93.5 ± 0.1	89.3 ± 0.2	96.2 ± 0.2	58.9 ± 0.3	75.8 ± 0.2	76.4 ± 0.1
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.3	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3 ± 0.1
16	VRF (Adam)	36.2 ± 0.7	39.9 ± 1.3	5.8 ± 0.5	34.8 ± 7.0	3.7 ± 0.4	60.7 ± 34.2	66.3 ± 13.0	55.9 ± 12.2	74.6 ± 10.7	34.2 ± 12.7	48.5 ± 7.4	41.9 ± 0.8
	VRF (SGD)	72.6 ± 0.1	75.1 ± 0.0	48.7 ± 0.4	92.1 ± 0.8	81.4 ± 0.1	87.2 ± 0.1	94.2 ± 0.2	96.5 ± 0.1	96.1 ± 0.1	71.0 ± 0.6	83.5 ± 0.2	81.7 ± 0.0
	WiSE (Adam)	46.5 ± 0.7	60.8 ± 0.2	2.4 ± 0.7	20.9 ± 8.0	2.6 ± 0.8	4.8 ± 4.4	3.8 ± 0.6	49.1 ± 20.1	79.3 ± 19.9	39.0 ± 13.7	49.5 ± 33.7	32.6 ± 4.3
	WiSE (SGD)	73.4 ± 0.1	73.6 ± 0.1	39.9 ± 0.3	83.4 ± 0.6	77.6 ± 0.2	87.5 ± 0.1	93.2 ± 0.2	86.0 ± 0.4	95.4 ± 0.0	63.5 ± 0.2	78.0 ± 0.2	77.4 ± 0.0
	CLIPood (Adam)	56.1 ± 0.1	65.0 ± 0.1	6.0 ± 1.5	40.0 ± 18.9	27.1 ± 30.1	24.1 ± 13.3	60.2 ± 27.9	81.5 ± 2.5	95.1 ± 0.2	56.6 ± 4.8	71.2 ± 11.3	53.0 ± 5.2
	CLIPood (SGD)	73.4 ± 0.2	73.0 ± 0.1	41.0 ± 0.6	91.8 ± 0.6	77.3 ± 0.2	86.8 ± 0.1	93.9 ± 0.1	94.4 ± 0.2	96.7 ± 0.1	65.1 ± 0.2	79.1 ± 0.1	79.3 ± 0.0
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.0	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6 ± 0.0

Table 32: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. OLM-SGD is compared with robust full-finetuning methods VRF, WiSE, and CLIPood, each trained with Adam and SGD. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	VRF (Adam)	23.1 ± 0.6	37.4 ± 3.5	10.7 ± 13.7	47.6 ± 13.9	26.0 ± 36.0	34.5 ± 2.7	38.0 ± 14.2	30.4 ± 9.2	71.7 ± 1.3	29.9 ± 0.5	43.5 ± 2.3	35.7 ± 5.3
	VRF (SGD)	76.7 ± 0.2	73.8 ± 0.0	40.0 ± 1.1	73.0 ± 1.0	80.4 ± 0.3	90.9 ± 0.1	94.4 ± 0.2	88.4 ± 0.4	96.0 ± 0.5	59.8 ± 0.8	81.0 ± 0.5	77.7 ± 0.1
	WiSE (Adam)	51.4 ± 0.7	61.8 ± 2.6	1.1 ± 0.1	18.2 ± 4.1	0.7 ± 0.1	56.8 ± 5.5	41.1 ± 12.7	16.3 ± 13.4	88.5 ± 0.7	39.9 ± 1.1	67.4 ± 1.4	40.3 ± 1.5
	WiSE (SGD)	76.1 ± 0.0	72.0 ± 0.1	37.2 ± 0.9	67.8 ± 1.1	79.8 ± 0.1	90.9 ± 0.1	94.3 ± 0.2	82.9 ± 0.4	95.6 ± 0.5	57.6 ± 0.7	78.4 ± 0.2	75.7 ± 0.1
	CLIPood (Adam)	58.1 ± 0.5	60.9 ± 0.4	1.0 ± 0.0	17.6 ± 6.1	13.4 ± 10.6	59.2 ± 2.7	73.9 ± 3.3	57.3 ± 10.5	90.2 ± 2.0	43.9 ± 0.9	68.4 ± 0.9	49.4 ± 2.3
	CLIPood (SGD)	76.2 ± 0.1	71.5 ± 0.2	37.5 ± 0.5	67.6 ± 0.6	78.8 ± 0.1	91.1 ± 0.1	94.0 ± 0.0	82.1 ± 0.3	95.6 ± 0.3	55.5 ± 0.4	77.0 ± 0.1	75.2 ± 0.2
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.6 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4 ± 0.1
2	VRF (Adam)	28.0 ± 1.4	38.0 ± 1.4	20.4 ± 13.7	33.2 ± 17.7	76.8 ± 0.1	31.9 ± 9.5	51.8 ± 30.6	43.8 ± 19.3	75.2 ± 0.6	32.8 ± 3.5	48.6 ± 5.8	43.7 ± 4.4
	VRF (SGD)	77.0 ± 0.2	75.7 ± 0.2	43.4 ± 1.8	79.0 ± 1.3	82.2 ± 0.3	91.1 ± 0.2	94.5 ± 0.0	92.7 ± 0.4	96.6 ± 0.2	65.2 ± 0.7	82.5 ± 0.7	80.0 ± 0.2
	WiSE (Adam)	54.3 ± 0.8	63.9 ± 1.8	1.1 ± 0.1	20.6 ± 1.5	0.7 ± 0.2	48.9 ± 10.5	28.1 ± 18.4	12.9 ± 9.8	89.5 ± 2.9	43.8 ± 5.5	65.9 ± 8.0	39.1 ± 0.7
	WiSE (SGD)	76.7 ± 0.0	73.3 ± 0.0	39.3 ± 0.7	71.1 ± 1.3	80.7 ± 0.2	91.1 ± 0.2	94.6 ± 0.1	85.7 ± 0.3	96.2 ± 0.3	61.1 ± 0.4	79.3 ± 0.5	77.2 ± 0.1
	CLIPood (Adam)	58.9 ± 0.5	65.5 ± 0.4	1.1 ± 0.1	27.9 ± 9.3	0.6 ± 0.1	64.0 ± 1.2	74.6 ± 1.2	78.3 ± 1.1	92.5 ± 0.3	50.5 ± 1.1	72.6 ± 1.7	53.3 ± 1.2
	CLIPood (SGD)	76.8 ± 0.0	72.7 ± 0.0	40.0 ± 0.2	71.0 ± 0.5	80.1 ± 0.1	91.2 ± 0.2	94.8 ± 0.2	86.0 ± 0.5	96.0 ± 0.1	58.2 ± 0.4	78.6 ± 0.2	76.9 ± 0.0
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1 ± 0.1
4	VRF (Adam)	32.0 ± 1.5	40.8 ± 1.8	23.9 ± 5.9	40.9 ± 15.7	37.8 ± 27.3	28.0 ± 38.1	43.7 ± 33.7	49.0 ± 4.4	80.8 ± 3.5	39.9 ± 4.3	53.4 ± 5.0	42.7 ± 3.9
	VRF (SGD)	77.9 ± 0.1	77.0 ± 0.2	49.0 ± 1.2	82.4 ± 2.3	83.9 ± 0.1	91.4 ± 0.1	95.1 ± 0.3	95.4 ± 0.2	97.4 ± 0.4	68.0 ± 0.7	84.7 ± 0.4	82.0 ± 0.2
	WiSE (Adam)	57.2 ± 0.3	66.3 ± 1.1	1.9 ± 0.4	18.7 ± 3.3	3.7 ± 3.3	1.5 ± 0.5	13.8 ± 14.6	39.1 ± 4.7	94.2 ± 0.6	54.9 ± 1.8	76.8 ± 0.3	38.9 ± 1.6
	WiSE (SGD)	77.5 ± 0.1	74.4 ± 0.2	43.3 ± 0.4	72.7 ± 1.4	82.4 ± 0.2	91.3 ± 0.1	94.9 ± 0.2	87.2 ± 0.5	96.7 ± 0.2	63.7 ± 0.2	81.2 ± 0.2	78.7 ± 0.2
	CLIPood (Adam)	61.6 ± 0.2	68.4 ± 0.5	2.5 ± 0.3	29.1 ± 17.2	18.0 ± 23.7	52.8 ± 27.3	80.9 ± 1.3	78.1 ± 8.3	95.6 ± 0.8	58.5 ± 0.8	79.0 ± 1.2	56.8 ± 5.5
	CLIPood (SGD)	77.5 ± 0.1	73.9 ± 0.1	42.9 ± 0.4	73.2 ± 0.3	81.7 ± 0.3	91.5 ± 0.0	95.3 ± 0.1	90.8 ± 0.3	96.5 ± 0.2	60.2 ± 0.3	80.5 ± 0.4	78.5 ± 0.1
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.3	91.3 ± 0.2	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.4	68.7 ± 0.9	84.9 ± 0.5	82.5 ± 0.2
8	VRF (Adam)	37.9 ± 0.7	40.1 ± 1.9	18.8 ± 6.8	50.1 ± 6.5	32.2 ± 29.5	51.1 ± 30.5	57.2 ± 7.6	70.4 ± 5.6	81.6 ± 1.6	44.2 ± 1.5	61.0 ± 3.2	49.5 ± 5.8
	VRF (SGD)	78.7 ± 0.2	78.4 ± 0.4	56.2 ± 0.6	88.8 ± 1.0	86.6 ± 0.2	91.8 ± 0.1	95.3 ± 0.2	96.9 ± 0.2	97.7 ± 0.2	71.8 ± 0.8	86.3 ± 0.2	84.4 ± 0.3
	WiSE (Adam)	58.5 ± 0.3	69.0 ± 0.4	2.3 ± 0.7	28.1 ± 3.0	5.2 ± 5.0	31.8 ± 27.6	45.8 ± 22.4	72.9 ± 11.9	94.7 ± 0.7	60.5 ± 2.1	81.3 ± 1.4	50.0 ± 5.0
	WiSE (SGD)	78.1 ± 0.1	75.7 ± 0.2	46.8 ± 0.5	79.8 ± 1.6	84.2 ± 0.1	91.6 ± 0.1	95.1 ± 0.1	89.4 ± 0.3	97.3 ± 0.1	66.3 ± 0.4	82.5 ± 0.1	80.6 ± 0.2
	CLIPood (Adam)	63.3 ± 0.1	71.0 ± 0.4	15.8 ± 8.8	23.3 ± 10.0	34.0 ± 30.2	60.8 ± 16.5	87.5 ± 1.2	91.5 ± 3.3	96.5 ± 0.4	63.5 ± 1.1	84.2 ± 0.3	62.8 ± 2.9
	CLIPood (SGD)	78.3 ± 0.0	75.5 ± 0.1	46.8 ± 0.6	81.7 ± 0.8	83.9 ± 0.2	91.7 ± 0.0	95.6 ± 0.1	94.9 ± 0.1	97.2 ± 0.3	63.3 ± 0.4	82.9 ± 0.1	81.1 ± 0.0
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7 ± 0.2
16	VRF (Adam)	42.9 ± 0.2	48.9 ± 0.9	16.4 ± 9.8	57.4 ± 17.0	52.1 ± 28.1	54.9 ± 4.7	68.8 ± 10.8	81.2 ± 5.5	87.2 ± 4.8	52.1 ± 4.0	62.1 ± 13.3	56.7 ± 2.6
	VRF (SGD)	79.8 ± 0.1	79.6 ± 0.1	62.4 ± 0.4	91.8 ± 0.9	88.1 ± 0.2	91.8 ± 0.1	95.9 ± 0.0	98.3 ± 0.2	97.9 ± 0.1	75.6 ± 0.5	87.8 ± 0.3	86.3 ± 0.2
	WiSE (Adam)	59.5 ± 0.0	69.5 ± 0.3	1.9 ± 0.6	30.5 ± 25.9	12.0 ± 13.1	50.4 ± 34.9	74.0 ± 4.3	86.7 ± 5.0	95.8 ± 0.6	60.8 ± 4.8	67.5 ± 22.6	55.3 ± 1.2
	WiSE (SGD)	78.9 ± 0.0	77.0 ± 0.0	51.8 ± 0.3	84.3 ± 0.2	85.7 ± 0.2	91.7 ± 0.1	95.5 ± 0.0	91.0 ± 0.1	97.3 ± 0.1	69.7 ± 0.1	84.2 ± 0.2	82.5 ± 0.0
	CLIPood (Adam)	65.3 ± 0.2	72.3 ± 0.0	29.4 ± 15.4	43.3 ± 19.3	34.6 ± 34.7	78.3 ± 0.9	89.3 ± 0.3	96.4 ± 1.3	97.2 ± 0.2	68.5 ± 1.4	86.6 ± 0.2	69.2 ± 5.2
	CLIPood (SGD)	79.6 ± 0.0	77.6 ± 0.1	53.4 ± 0.5	88.8 ± 0.5	86.2 ± 0.2	91.9 ± 0.1	96.1 ± 0.1	97.9 ± 0.1	98.0 ± 0.1	68.0 ± 0.2	85.9 ± 0.2	83.9 ± 0.1
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8 ± 0.1

H CAUSAL TRAJECTORY ABLATIONS

This appendix examines the optimizer choice and the components of OLM-SGD. Across the comparisons, we use four complementary views: fixed-schedule few-shot accuracy, cross-dataset transfer of the selected training epoch, long-horizon dynamics, and best-to-final accuracy drop. The fixed-schedule results measure adaptation quality under the stated schedules. The trajectory experiments expose sensitivity that a single endpoint can conceal. Appendix H.5 focuses on fixed-schedule few-shot accuracy.

Shared setup. We evaluate the optimizer variants considered in this ablation study on 11 datasets: ImageNet, Caltech101, Oxford Pets, Stanford Cars, Flowers102, Food101, FGVC-Aircraft, SUN397, DTD, EuroSAT, and UCF101. Each experiment uses seeds 1, 2, and 3. All optimizers adapt the same Transformer-block parameters in both CLIP encoders. They use the supervised classification objective in Appendix A. The prompt strings remain fixed. The algorithms and optimizer-specific settings are given in Appendix C and Appendix D.

Fixed-schedule and long-horizon evaluations. Unless stated otherwise, the few-shot tables cover ViT-B/32, ViT-B/16, and ViT-L/14 with $K \in \{1, 2, 4, 8, 16\}$ labeled examples per class. Adam, randomized SGD, and GaLoRE use 30 epochs at 10^{-4}. CTF-SGD uses randomized minibatches and 30 epochs at 10^{-3}. The class-unique SGD results in Appendix H.2 use 30 epochs at 10^{-3}. OLM-SGD, its class-mean variant, and its variant without polar reconstruction use 70 epochs at 10^{-4}. The trajectory experiments instead use a common horizon of 300 epochs for 1-shot CLIP (ViT-B/16) adaptation. They evaluate learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Other optimizer settings remain unchanged. Randomized minibatches have capacity 64. OLM-SGD uses class-unique minibatches with capacity $B = \min\{64, C\}$. At $K = 1$, every support label is unique, so the two samplers have the same batching distribution and update count per epoch. Their realized random orders need not coincide.

Aggregation and epoch transfer. Accuracy entries and trajectory bands report the mean and population standard deviation across the three seeds. Dataset averages give equal weight to all 11 datasets unless a subsection specifies a smaller common set. For an aggregate mean and standard deviation, we first average datasets within each seed. The epoch-transfer diagnostic uses learning rate 10^{-4}. Ties within 10^{-12} percentage points select the earliest epoch. For each method and source dataset, it selects the earliest epoch in 1–300 that maximizes the source's seed-mean *test* accuracy. Each independently trained target is evaluated at exactly that epoch. Only the duration is transferred, not weights or examples. Heatmap entries subtract the target's zero-shot CLIP accuracy. The diagonal uses each target's own best epoch. Transfer summaries exclude that diagonal and average the 110 ordered pairs of distinct datasets when all 11 datasets are included. The oracle gap is the target's best seed-mean accuracy minus its accuracy at the transferred epoch. This is a test-guided diagnostic, not validation-selected early stopping.

Accuracy retention and loss coverage. For each dataset and seed, the best-to-final drop is the highest test accuracy over epochs 1–300 minus the accuracy at epoch 300. We compute this difference before averaging seeds. This differs from maximizing the seed-mean curve in the transfer diagnostic. The zero-shot checkpoint is not included in the maximization. Accuracy changes and drops are expressed in percentage points. Loss curves show the recorded training loss, not test loss. A loss mean and band are shown only where all three seeds have finite values. Nonfinite losses are not interpolated, and they do not remove otherwise complete test-accuracy histories.

H.1 Comparison against Adam and SGD

Adam tests whether adaptive moment estimation is sufficient to stabilize few-shot full fine-tuning. SGD provides a stronger first-order baseline based on the arithmetic mean of the gradients within each randomized minibatch. Both methods aggregate the minibatch using the standard mean gradient. In contrast, OLM-SGD keeps the individual example gradients separate and replaces their direct arithmetic averaging with a structured tensor-wise merge. For each trainable tensor, it retains the dominant rank-r components of each individual gradient and combines them through polar reconstruction. All 396 trajectories for these three methods contain complete test-accuracy histories through epoch 300. The following analyses use the shared protocol above.

H.1.1 Few-Shot Accuracy across Backbones and Shot Levels

Average accuracy. Tables 34–36 report the detailed results. Table 33 summarizes their Average columns. OLM-SGD exceeds SGD in 14 of the 15 backbone–shot settings. On ViT-B/32, its advantage ranges from 0.3 to 0.9 points. On ViT-B/16, it ranges from 0.5 to 1.2 points. Averaging the five displayed margins gives 0.70 and 0.86 points, respectively. The improvement is smaller on ViT-L/14, where the mean margin is 0.18 points. In particular, SGD leads at 1 shot with 77.8%, compared with 77.4% for OLM-SGD. OLM-SGD then leads by 0.1–0.5 points at 2–16 shots. Thus, the overall advantage persists across model scales but is not equally large.

Dataset-level gains and exceptions. Using the displayed means, OLM-SGD strictly exceeds SGD in $42/55$, $45/55$, and $29/55$ dataset–shot settings for ViT-B/32, ViT-B/16, and ViT-L/14. There are three, two, and three additional ties, respectively. The advantage is particularly consistent on Stanford Cars, where OLM-SGD leads in all 15 settings. It also leads in $14/15$ settings on Flowers102 and $13/15$ on each of SUN397, EuroSAT, and DTD. For example, 16-shot ViT-B/16 accuracy on Aircraft improves from 48.6% with SGD to 52.9% with OLM-SGD. At 8 shots on EuroSAT, the corresponding values are 88.3% and 90.8%. These repeated gains show that the average improvement is not confined to one dataset.

ImageNet and Caltech101 provide important counterexamples. SGD exceeds OLM-SGD in $12/15$ ImageNet settings and $9/15$ Caltech101 settings. On ImageNet with 16-shot ViT-B/16, SGD reaches 72.1%, whereas OLM-SGD obtains 70.8%. The fixed-schedule tables support an average advantage, not dominance on every task. They also combine different epoch budgets and, when $K > 1$, different batching rules. The common-horizon 1-shot comparisons below provide a more controlled assessment of optimizer behavior.

Table 33: Dataset-average few-shot accuracy (%) from Tables 34–36. We reproduce the displayed Average means. The final column subtracts SGD from OLM-SGD and is measured in percentage points. All optimizers use $\eta = 10^{-4}$. Bold marks the highest displayed average in each row.

Backbone	Shots	Adam	SGD	OLM-SGD	OLM-SGD − SGD
ViT-B/32	1	13.2	66.9	**67.2**	+0.3
	2	14.5	69.8	**70.4**	+0.6
	4	18.7	72.3	**73.1**	+0.8
	8	22.5	74.9	**75.8**	+0.9
	16	30.3	77.6	**78.5**	+0.9
ViT-B/16	1	12.4	71.5	**72.0**	+0.5
	2	12.5	74.5	**75.1**	+0.6
	4	19.5	76.8	**77.8**	+1.0
	8	22.9	79.1	**80.3**	+1.2
	16	27.6	81.6	**82.6**	+1.0
ViT-L/14	1	23.7	**77.8**	77.4	−0.4
	2	23.7	80.0	**80.1**	+0.1
	4	28.3	82.1	**82.5**	+0.4
	8	36.8	84.4	**84.7**	+0.3
	16	49.5	86.3	**86.8**	+0.5

Adam is not a sufficient baseline for full fine-tuning. Adam is below both SGD and OLM-SGD in every reported backbone–shot average. At 1 shot, its averages are 13.2%, 12.4%, and 23.7% on the three backbones. SGD instead reaches 66.9%, 71.5%, and 77.8%. Thus, poor Adam results under this schedule do not imply that full fine-tuning is intrinsically ineffective. SGD already performs strongly. The smaller additional gains over SGD are consequently more informative than the much larger gaps to Adam.

Table 34: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. OLM-SGD is compared with SGD and Adam. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	SGD	61.2 ± 0.3	67.0 ± 0.2	22.3 ± 0.5	62.6 ± 5.5	61.4 ± 0.6	80.3 ± 0.2	89.1 ± 0.6	77.7 ± 0.8	93.6 ± 0.6	49.7 ± 0.3	70.4 ± 0.5	66.9 ± 0.4
	Adam	6.8 ± 1.6	20.1 ± 1.7	1.1 ± 0.1	11.7 ± 2.3	0.5 ± 0.0	2.8 ± 2.1	8.3 ± 3.0	3.6 ± 2.6	52.0 ± 1.7	16.3 ± 1.9	22.3 ± 1.7	13.2 ± 0.2
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.6	50.7 ± 1.9	70.1 ± 0.4	67.2 ± 0.5
2	SGD	61.6 ± 0.2	68.2 ± 0.3	23.9 ± 0.2	75.3 ± 2.2	63.3 ± 0.1	80.6 ± 0.1	89.8 ± 0.5	82.6 ± 0.8	94.5 ± 0.1	54.6 ± 1.6	72.9 ± 0.4	69.8 ± 0.1
	Adam	9.5 ± 0.7	17.7 ± 1.1	1.0 ± 0.0	16.3 ± 8.2	0.5 ± 0.1	2.3 ± 1.9	7.1 ± 4.1	2.4 ± 1.7	64.1 ± 3.3	18.5 ± 4.8	20.9 ± 3.4	14.5 ± 0.1
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.9 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.4 ± 0.9	70.4 ± 0.1
4	SGD	62.9 ± 0.3	69.6 ± 0.3	27.5 ± 0.3	81.2 ± 1.3	66.3 ± 0.5	80.8 ± 0.2	90.6 ± 0.4	87.2 ± 0.8	94.8 ± 0.1	58.4 ± 0.5	75.6 ± 0.7	72.3 ± 0.2
	Adam	13.5 ± 0.7	21.3 ± 0.5	1.6 ± 0.3	22.8 ± 9.6	1.1 ± 0.3	2.3 ± 0.7	15.6 ± 8.6	16.0 ± 11.6	62.9 ± 7.0	24.8 ± 3.3	24.5 ± 7.0	18.7 ± 1.4
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.1 ± 0.6	81.4 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1 ± 0.2
8	SGD	64.0 ± 0.2	71.0 ± 0.2	32.4 ± 0.7	88.5 ± 0.7	70.1 ± 0.3	81.0 ± 0.2	90.9 ± 0.4	90.5 ± 0.6	95.2 ± 0.4	62.4 ± 0.4	77.8 ± 0.1	74.9 ± 0.2
	Adam	17.9 ± 0.3	27.0 ± 0.6	1.7 ± 0.1	18.5 ± 5.5	1.4 ± 0.5	3.9 ± 2.4	7.3 ± 3.0	35.9 ± 13.2	70.9 ± 6.7	27.9 ± 5.5	35.5 ± 2.4	22.5 ± 0.2
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.6 ± 0.1	91.2 ± 0.5	91.9 ± 0.1	95.2 ± 0.2	64.6 ± 0.6	78.5 ± 0.4	75.8 ± 0.1
16	SGD	66.6 ± 0.2	72.3 ± 0.1	39.5 ± 0.3	91.9 ± 0.4	74.5 ± 0.3	81.3 ± 0.0	90.9 ± 0.5	93.8 ± 0.1	95.5 ± 0.1	67.1 ± 0.6	79.9 ± 0.1	77.6 ± 0.1
	Adam	23.0 ± 0.4	30.4 ± 1.0	2.0 ± 0.6	46.3 ± 16.9	3.1 ± 1.5	8.2 ± 2.9	21.0 ± 2.3	51.5 ± 15.0	77.7 ± 3.2	38.1 ± 2.3	31.4 ± 15.5	30.3 ± 1.3
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5 ± 0.1

H.1.2 Cross-Dataset Transfer of the Selected Training Epoch

Retaining accuracy under a transferred duration. Figure 9 and Table 37 evaluate the source-selected stopping times. OLM-SGD achieves 71.73% mean transfer accuracy, compared with 70.64% for SGD and 40.65% for Adam. Its mean gap to the target oracle is 0.91 points. The corresponding gaps are 1.50 for SGD and 9.35 for Adam. OLM-SGD remains within one point of the target oracle on $85/110$ pairs, compared with $75/110$ for SGD.

The transfer advantage over SGD is not explained only by a higher attainable peak. OLM-SGD's target-oracle average is 0.50 points higher, whereas its transfer average is 1.10 points higher. The remaining 0.60 points come from the smaller penalty for transferring the duration. Its largest oracle

Table 35: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. OLM-SGD is compared with SGD and Adam. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	SGD	67.5 ± 0.2	69.6 ± 0.3	29.4 ± 0.2	68.1 ± 0.6	69.4 ± 0.9	85.9 ± 0.6	92.6 ± 0.4	82.5 ± 0.4	94.2 ± 0.1	53.2 ± 0.7	74.2 ± 0.1	71.5 ± 0.1
	Adam	8.4 ± 0.6	19.4 ± 3.8	1.2 ± 0.3	13.2 ± 2.8	0.5 ± 0.1	1.0 ± 0.0	4.2 ± 1.9	0.9 ± 0.2	50.0 ± 5.6	19.1 ± 2.0	18.0 ± 5.9	12.4 ± 0.4
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.3	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0 ± 0.3
2	SGD	67.9 ± 0.2	71.4 ± 0.1	32.3 ± 1.2	77.9 ± 1.1	71.7 ± 0.4	86.0 ± 0.1	92.8 ± 0.4	88.6 ± 0.6	94.5 ± 0.3	59.5 ± 1.0	77.3 ± 0.3	74.5 ± 0.2
	Adam	9.2 ± 5.4	22.1 ± 3.7	1.0 ± 0.0	13.1 ± 5.4	0.5 ± 0.0	1.0 ± 0.0	2.8 ± 0.1	5.3 ± 5.5	64.0 ± 5.5	14.1 ± 6.8	4.7 ± 5.2	12.5 ± 1.4
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.2 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1 ± 0.2
4	SGD	68.9 ± 0.2	72.4 ± 0.2	36.3 ± 0.6	82.6 ± 1.5	74.7 ± 0.3	86.5 ± 0.2	93.4 ± 0.2	91.9 ± 0.5	95.3 ± 0.1	63.1 ± 0.4	79.6 ± 0.2	76.8 ± 0.1
	Adam	16.6 ± 0.8	27.2 ± 0.7	1.2 ± 0.2	15.8 ± 4.0	0.8 ± 0.1	3.2 ± 3.2	13.2 ± 14.0	13.4 ± 7.1	75.6 ± 1.2	18.7 ± 1.3	28.8 ± 9.4	19.5 ± 1.0
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.0	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8 ± 0.1
8	SGD	70.1 ± 0.3	73.7 ± 0.3	42.2 ± 0.6	88.3 ± 0.7	77.7 ± 0.4	86.8 ± 0.3	93.5 ± 0.3	94.1 ± 0.2	95.6 ± 0.1	66.3 ± 0.9	81.5 ± 0.3	79.1 ± 0.3
	Adam	22.8 ± 0.2	32.8 ± 1.5	1.9 ± 0.1	13.1 ± 3.1	0.9 ± 0.1	1.2 ± 0.2	6.9 ± 4.3	32.7 ± 10.5	69.8 ± 13.7	33.7 ± 2.4	35.9 ± 14.0	22.9 ± 2.5
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.3	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3 ± 0.1
16	SGD	72.1 ± 0.1	74.8 ± 0.1	48.6 ± 0.3	92.1 ± 0.8	81.3 ± 0.1	87.0 ± 0.1	94.0 ± 0.1	96.5 ± 0.1	96.3 ± 0.2	71.0 ± 0.4	83.7 ± 0.2	81.6 ± 0.0
	Adam	29.4 ± 0.7	37.7 ± 1.1	3.1 ± 0.8	26.0 ± 4.2	2.9 ± 0.6	4.7 ± 3.6	3.5 ± 0.6	54.6 ± 12.0	72.5 ± 11.2	33.4 ± 12.7	36.0 ± 22.6	27.6 ± 2.5
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.0	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6 ± 0.0

Table 36: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. OLM-SGD is compared with SGD and Adam. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average gives equal weight to the 11 datasets within each seed, then reports the mean and standard deviation over seeds.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	SGD	76.6 ± 0.1	73.9 ± 0.0	40.0 ± 1.4	73.5 ± 1.0	80.3 ± 0.3	90.9 ± 0.1	94.4 ± 0.2	88.7 ± 0.4	96.1 ± 0.6	60.1 ± 0.9	81.4 ± 0.6	77.8 ± 0.1
	Adam	14.1 ± 1.0	32.9 ± 4.7	1.0 ± 0.1	18.6 ± 2.4	0.9 ± 0.5	25.8 ± 3.8	25.3 ± 10.9	18.6 ± 13.1	61.1 ± 1.0	25.1 ± 0.7	37.4 ± 3.8	23.7 ± 1.6
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.6 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4 ± 0.1
2	SGD	76.8 ± 0.3	75.8 ± 0.3	43.4 ± 1.7	79.4 ± 1.6	82.1 ± 0.3	91.1 ± 0.2	94.5 ± 0.0	92.9 ± 0.4	96.7 ± 0.2	65.3 ± 0.6	82.5 ± 0.8	80.0 ± 0.3
	Adam	19.0 ± 1.3	34.6 ± 2.5	1.0 ± 0.0	13.7 ± 2.9	0.8 ± 0.1	15.0 ± 10.5	18.6 ± 10.7	19.2 ± 11.0	70.7 ± 8.6	27.8 ± 2.7	39.9 ± 8.7	23.7 ± 1.5
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1 ± 0.1
4	SGD	77.7 ± 0.1	77.0 ± 0.2	49.0 ± 1.3	83.1 ± 2.1	83.9 ± 0.2	91.4 ± 0.1	95.1 ± 0.3	95.6 ± 0.3	97.6 ± 0.4	68.2 ± 0.9	84.8 ± 0.4	82.1 ± 0.2
	Adam	23.0 ± 0.3	35.8 ± 2.7	1.9 ± 0.4	11.2 ± 6.2	3.6 ± 2.9	1.7 ± 1.0	14.0 ± 14.3	48.0 ± 1.3	82.8 ± 1.9	37.6 ± 3.8	52.0 ± 2.8	28.3 ± 2.5
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.3	91.3 ± 0.2	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.4	68.7 ± 0.9	84.9 ± 0.5	82.5 ± 0.2
8	SGD	78.4 ± 0.2	78.5 ± 0.5	56.4 ± 0.6	89.0 ± 1.0	86.5 ± 0.3	91.7 ± 0.1	95.2 ± 0.2	97.0 ± 0.1	97.7 ± 0.3	72.0 ± 0.6	86.3 ± 0.2	84.4 ± 0.3
	Adam	28.6 ± 0.9	38.5 ± 0.7	2.5 ± 0.6	24.4 ± 10.6	5.9 ± 6.5	15.5 ± 17.1	37.6 ± 24.7	68.7 ± 6.9	80.4 ± 2.6	44.9 ± 2.9	57.7 ± 2.0	36.8 ± 4.7
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7 ± 0.2
16	SGD	79.4 ± 0.1	79.6 ± 0.1	62.6 ± 0.2	91.8 ± 0.8	88.0 ± 0.3	91.8 ± 0.1	95.8 ± 0.1	98.3 ± 0.1	97.9 ± 0.1	75.7 ± 0.4	88.0 ± 0.2	86.3 ± 0.2
	Adam	35.4 ± 0.5	42.6 ± 2.4	4.1 ± 1.2	41.8 ± 8.1	21.2 ± 11.4	47.7 ± 2.4	69.7 ± 7.6	85.1 ± 2.8	89.1 ± 0.6	52.9 ± 0.7	55.2 ± 13.4	49.5 ± 1.2
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8 ± 0.1

gap is also smaller: 7.17 points rather than 26.15. The largest SGD gap occurs when Food101 selects epoch 4 for EuroSAT. This illustrates the cost of transferring an early stopping time to a target that benefits from longer training.

Table 37: Epoch-transfer diagnostic for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. The target oracle averages each dataset's best seed-mean accuracy. Transfer accuracy and oracle gaps average the 110 off-diagonal source–target pairs. The final column counts pairs within one percentage point of the target oracle. These are test-guided diagnostics.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp
Adam	49.99	40.65	9.35	75/110
SGD	72.14	70.64	1.50	75/110
OLM-SGD	**72.64**	**71.73**	**0.91**	**85/110**

Different optimal epochs can still transfer well. OLM-SGD's selected epochs range from 18 on ImageNet to 292 on EuroSAT. SGD's range is 4–270. Thus, the improvement does not require the task-specific maxima to occur at similar epochs. Instead, OLM-SGD retains more of the target's attainable accuracy across the source-selected durations. Adam selects epoch 1 on nine datasets. Its apparent proximity to its own oracle on many pairs is not evidence of successful adaptation, because that oracle is already weak. Absolute transfer accuracy is needed to interpret the gap.

ImageNet remains the main limitation. OLM-SGD improves over zero-shot CLIP on $103/110$ off-diagonal pairs. All seven negative entries target ImageNet. The largest decrease is 4.11 points, compared with 9.29 points for SGD. All 100 transfers to the other ten targets improve over zero-shot CLIP for OLM-SGD. Nevertheless, SGD has one more positive entry overall, with $104/110$. Hence, OLM-SGD's advantage is better accuracy retention and smaller transfer losses, not a uniformly larger count of positive cells. A late stopping time can still harm ImageNet, even with OLM-SGD.

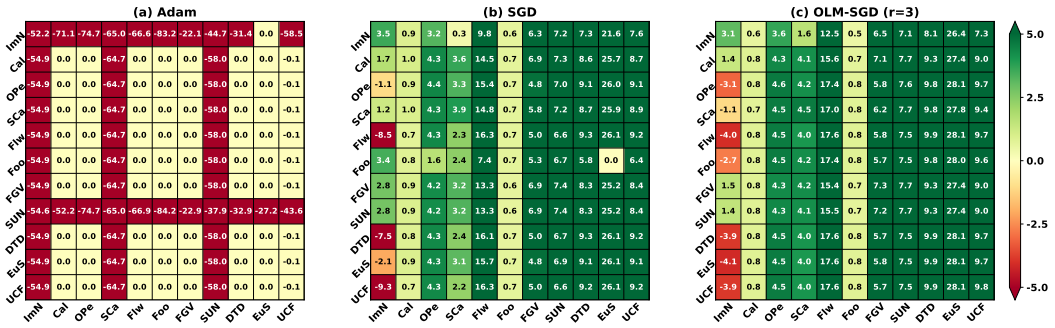

Figure 9: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Each source row selects the earliest epoch in 1–300 that maximizes mean *test* accuracy across three seeds. Each target column evaluates its independently trained trajectory at that epoch. Entries give the target accuracy change relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Colors saturate at ± 5 points, while the annotations retain the rounded numerical values. OLM-SGD uses $r = 3$. Only the duration is transferred. Diagonal entries use target-specific oracle selection. This is a test-guided diagnostic, not validation-selected early stopping.

H.1.3 Long-Horizon Dynamics and Learning-Rate Sensitivity

A broader useful learning-rate range. Figures 10–13 show the test-accuracy trajectories. Table 38 summarizes their endpoints. At 10^{-3}, OLM-SGD reaches 69.23% mean final accuracy, compared with 58.67% for SGD and 2.36% for Adam. OLM-SGD exceeds SGD on all 11 datasets. Large gains occur on Stanford Cars (53.31% versus 16.94%), Aircraft (29.36% versus 5.10%), and Flowers102 (86.26% versus 63.46%). At 10^{-4}, OLM-SGD again leads on every dataset. The mean final accuracies are 71.75%, 70.46%, and 15.37% for OLM-SGD, SGD, and Adam.

Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of mean final accuracy is 2.52 points for OLM-SGD. It is 12.79 for SGD and 58.96 for Adam. Including 10^{-2} expands these ranges to 24.37, 59.99, and 59.20 points. Thus, OLM-SGD is less sensitive to the learning rate.

Table 38: Long-horizon summary for 1-shot CLIP (ViT-B/16). Final accuracy is evaluated at epoch 300. Best-to-final drops are computed within each seed before aggregation. For both quantities, we first average the 11 datasets within each seed and then report mean $\pm$ population standard deviation across seeds. The last column gives the mean accuracy change from epoch 30 to epoch 300. All 396 complete accuracy histories are included.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	Adam	2.12 ± 0.47	44.36 ± 0.47	$+0.35$
	SGD	11.47 ± 0.88	37.84 ± 0.77	$+7.60$
	OLM-SGD	47.38 ± 0.43	3.45 ± 0.80	-0.18
10^{-3}	Adam	2.36 ± 0.51	44.13 ± 0.51	$+0.19$
	SGD	58.67 ± 2.09	6.49 ± 1.35	$+0.60$
	OLM-SGD	69.23 ± 0.54	2.22 ± 0.07	-0.91
10^{-4}	Adam	15.37 ± 1.39	34.76 ± 1.87	$+3.46$
	SGD	70.46 ± 0.23	1.81 ± 0.12	-1.03
	OLM-SGD	71.75 ± 0.26	0.95 ± 0.06	$+0.14$
10^{-5}	Adam	61.33 ± 0.63	7.52 ± 0.86	-0.53
	SGD	71.46 ± 0.02	0.56 ± 0.05	$+1.49$
	OLM-SGD	71.75 ± 0.26	0.11 ± 0.00	$+3.85$

Prolonged training and the smaller-rate trade-off. At 10^{-4}, OLM-SGD gains 0.14 points from epoch 30 to epoch 300, whereas SGD loses 1.03. From epoch 70, both decline, but OLM-SGD loses 0.27 points rather than 0.72. This is reduced degradation, not monotonic improvement. Seed variability is not uniformly lower either. Averaging the dataset-specific per-epoch standard deviations gives 0.57 points for OLM-SGD and 0.43 for SGD.

At 10^{-5}, SGD reaches 71.46% final accuracy, only 0.29 points below OLM-SGD, and leads on six datasets. OLM-SGD adapts more slowly: its epoch-30 average is 67.89%, compared with 69.97% for SGD. Continuing to epoch 300 adds 3.85 and 1.49 points, respectively. Adam also recovers to 61.33%. Thus, smaller rates improve retention but can delay adaptation.

Lower training loss need not yield better test accuracy. Figures 14–17 separate support fitting from generalization. At 10^{-4}, SGD has lower mean final training loss than OLM-SGD on every dataset, yet lower final test accuracy on every dataset. On Food101, Adam's final loss is approximately 2.32×10^{-4}, but test accuracy is only 2.22%. OLM-SGD retains 86.02% accuracy with a larger loss of approximately 5.48×10^{-3}. Therefore, lower support loss is not sufficient for effective adaptation.

Numerical failure is distinct from poor generalization. At 10^{-2}, Adam's SUN397 losses become nonfinite from epochs 185, 231, and 235 for seeds 1, 2, and 3. Its complete test-accuracy histories remain included. The missing loss segments do not indicate early stopping. OLM-SGD avoids these nonfinite losses, but this comparison alone cannot attribute that benefit to a particular merge component.

H.1.4 Best-to-Final Test-Accuracy Drop

Smaller mean degradation at every tested rate. Tables 39–42 quantify the retention visible in the accuracy curves. OLM-SGD has the smallest dataset-average drop at all four learning rates. Its drops are 3.45, 2.22, 0.95, and 0.11 points at 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The corresponding values are 37.84, 6.49, 1.81, and 0.56 for SGD, and 44.36, 44.13, 34.76, and 7.52 for Adam. Consequently, the improvement remains visible even when SGD becomes competitive in absolute accuracy.

Dataset-level retention and the ImageNet outlier. At 10^{-4}, OLM-SGD has a smaller mean drop than SGD on every dataset. It remains within one point of its seed-specific peak on nine datasets,

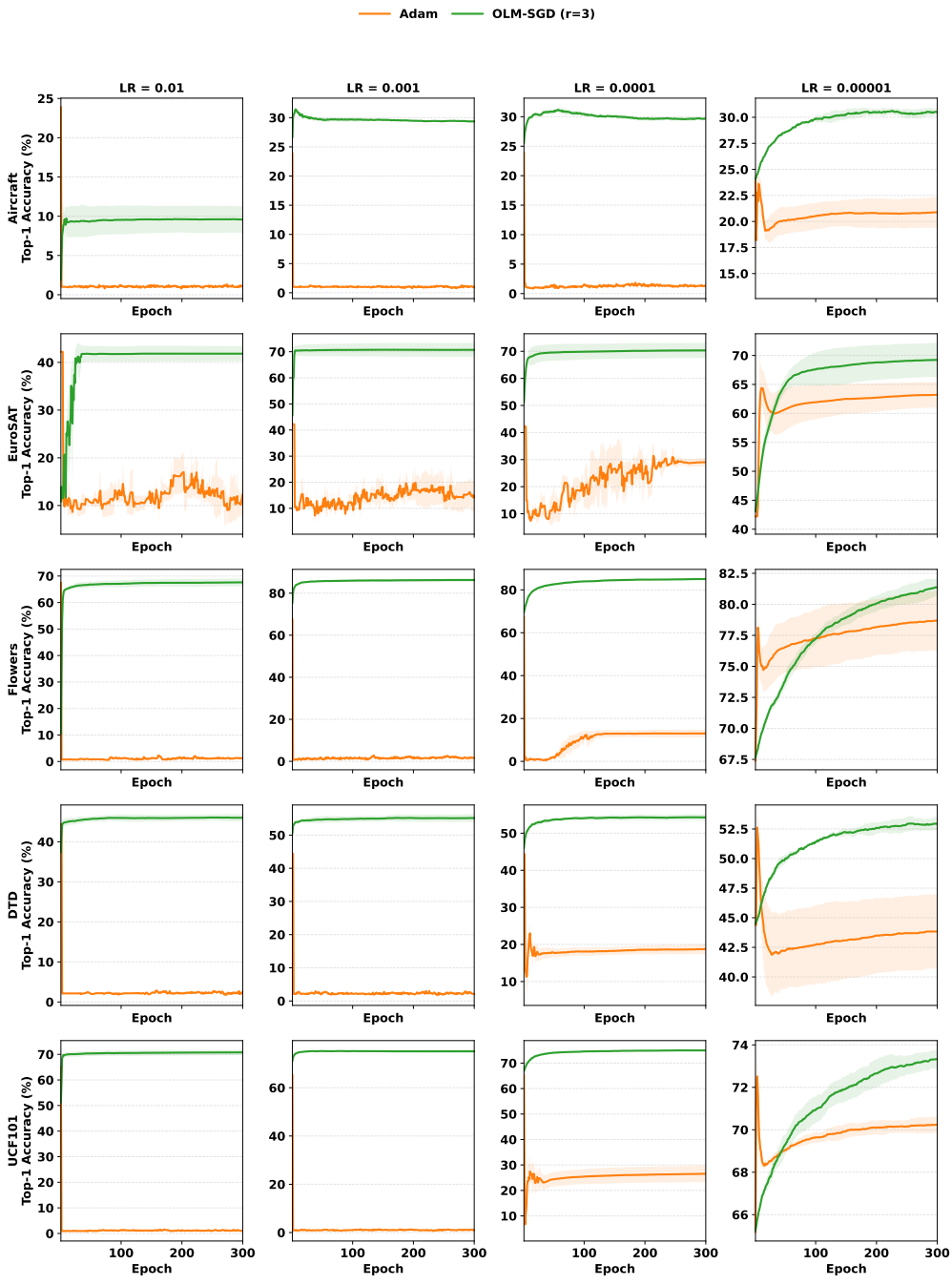

Figure 10: Test-accuracy trajectories for Adam and OLM-SGD ($r = 3$). Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The shared protocol uses 1-shot CLIP (ViT-B/16) over 300 epochs. Lines show means and shaded bands show one population standard deviation across seeds 1, 2, and 3.

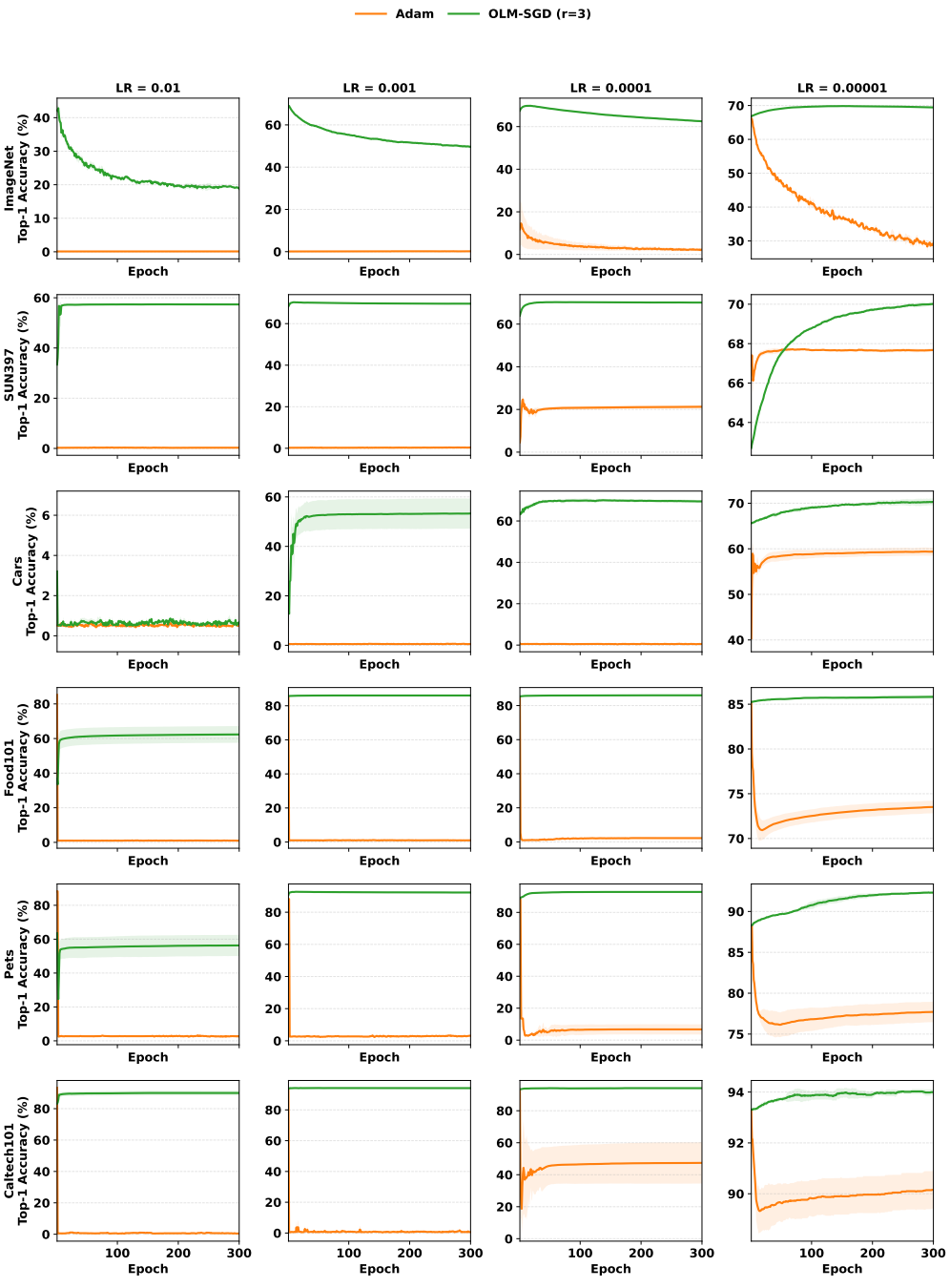

Figure 11: Test-accuracy trajectories for Adam and OLM-SGD ($r = 3$), continuing Figure 10. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. ImageNet retains training augmentation under the shared protocol.

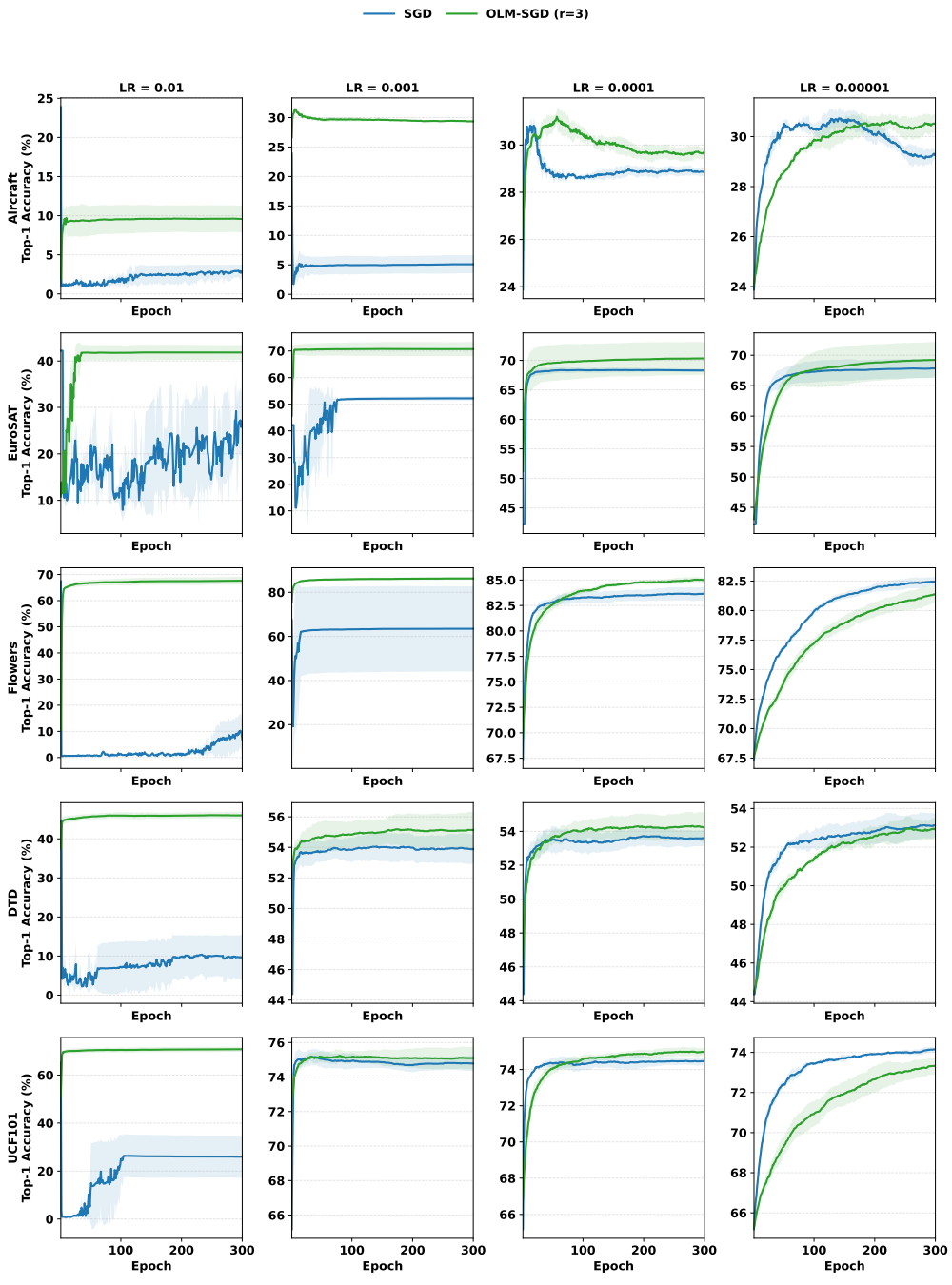

Figure 12: Test-accuracy trajectories for SGD and OLM-SGD ($r = 3$). Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The shared protocol uses 1-shot CLIP (ViT-B/16) over 300 epochs. Lines show means and shaded bands show one population standard deviation across seeds 1, 2, and 3.

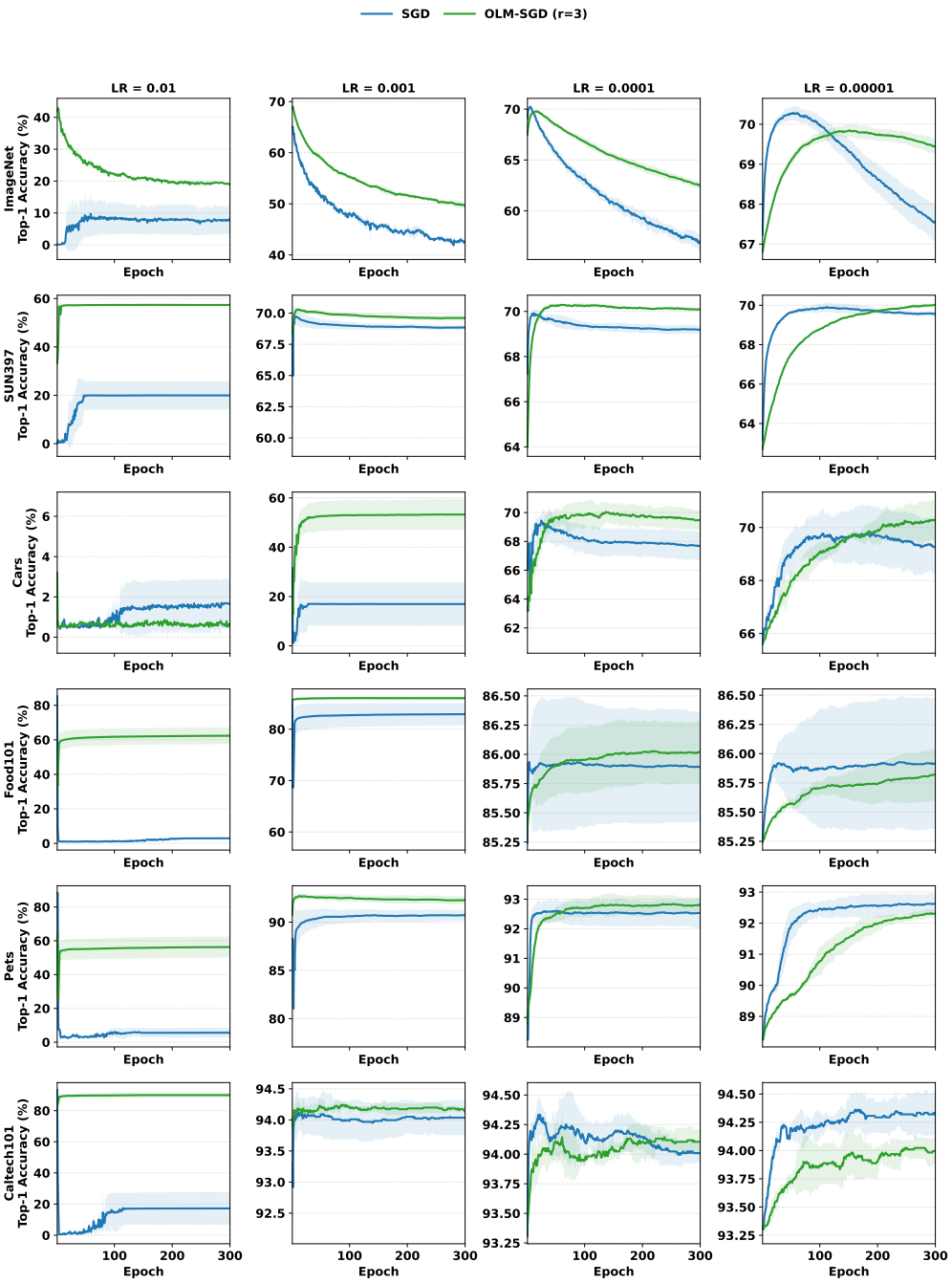

Figure 13: Test-accuracy trajectories for SGD and OLM-SGD ($r = 3$), continuing Figure 12. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. ImageNet retains training augmentation under the shared protocol.

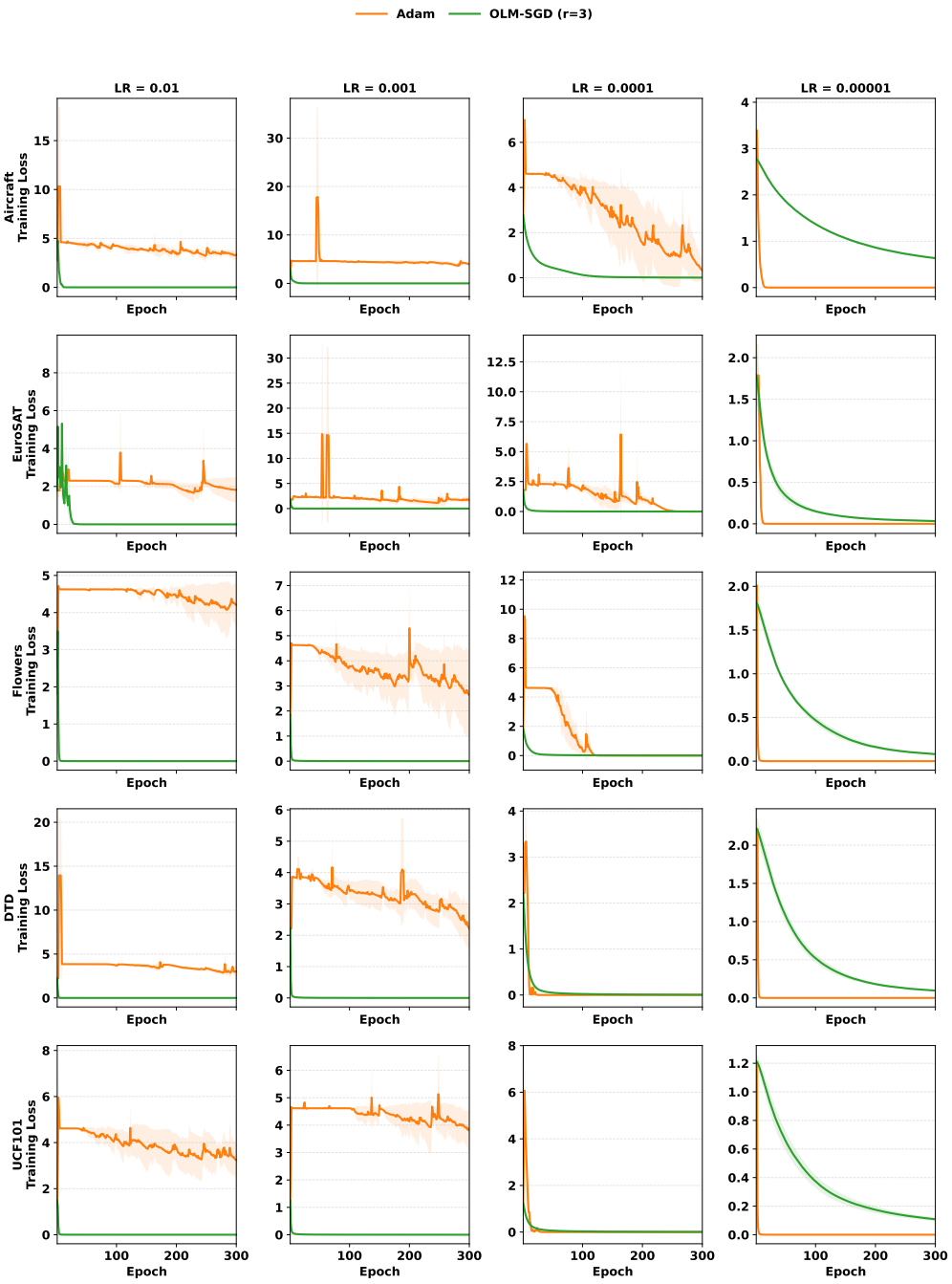

Figure 14: Training loss for the Adam and OLM-SGD trajectories in Figure 10. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across the three seeds. These are support-set training losses, not test losses.

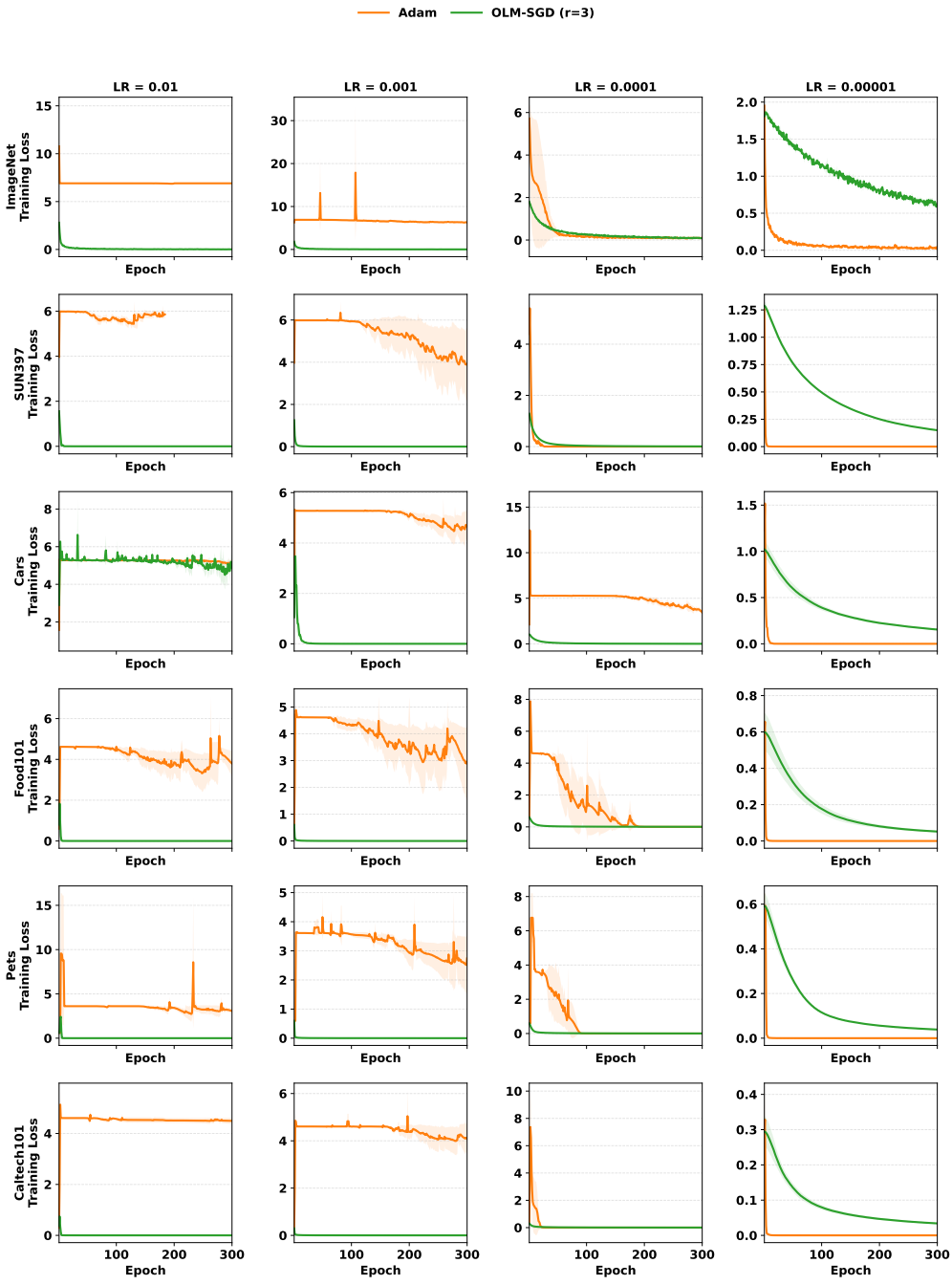

Figure 15: Training loss for the Adam and OLM-SGD trajectories in Figure 11. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Curves use epochs with finite losses in all three seeds. At 10^{-2}, Adam's SUN397 losses become nonfinite from epochs 185, 231, and 235 for seeds 1, 2, and 3. The corresponding loss mean and band therefore end after epoch 184. All test-accuracy histories still contain 300 epochs. No loss values are extrapolated.

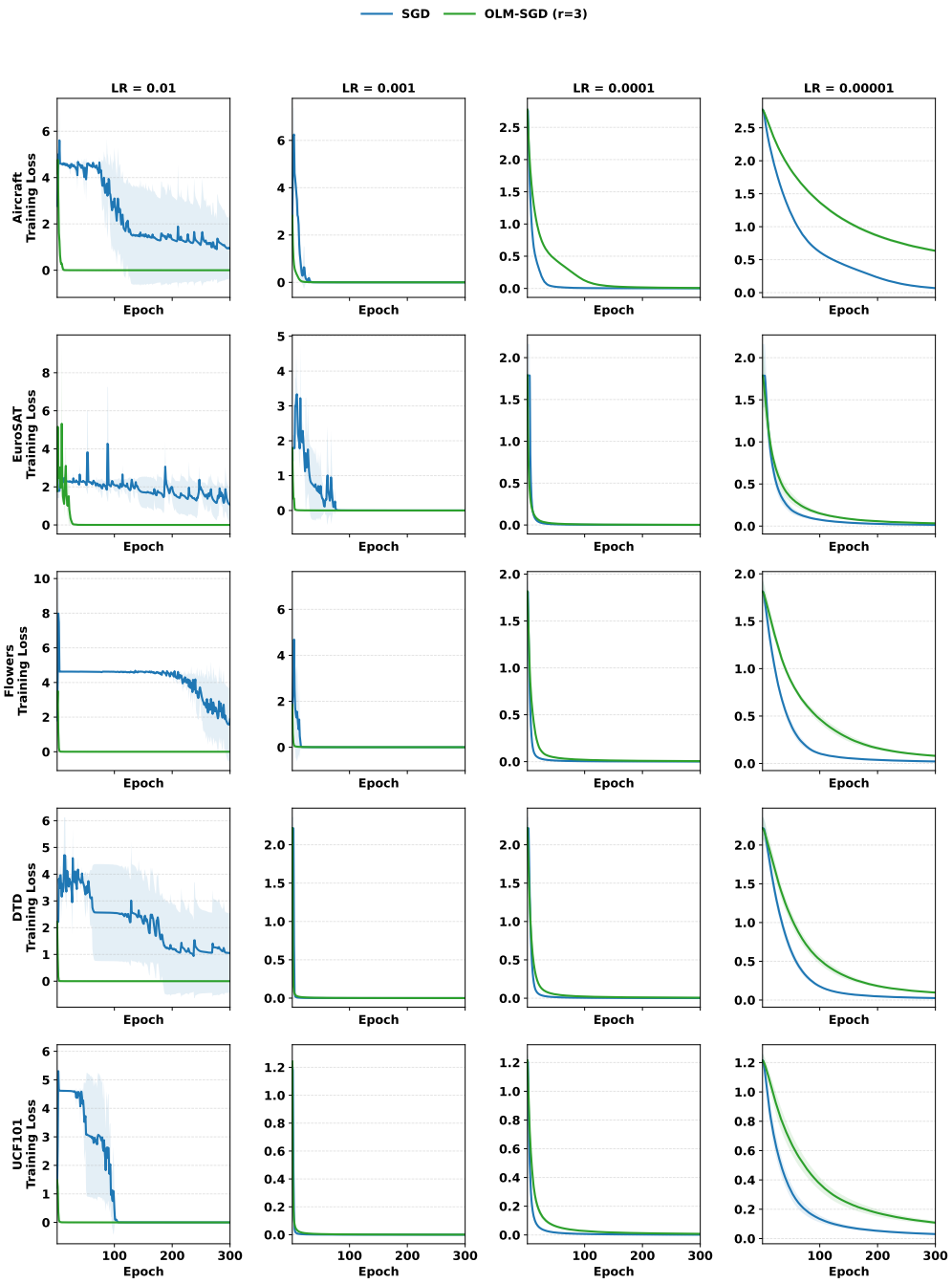

Figure 16: Training loss for the SGD and OLM-SGD trajectories in Figure 12. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across the three seeds.

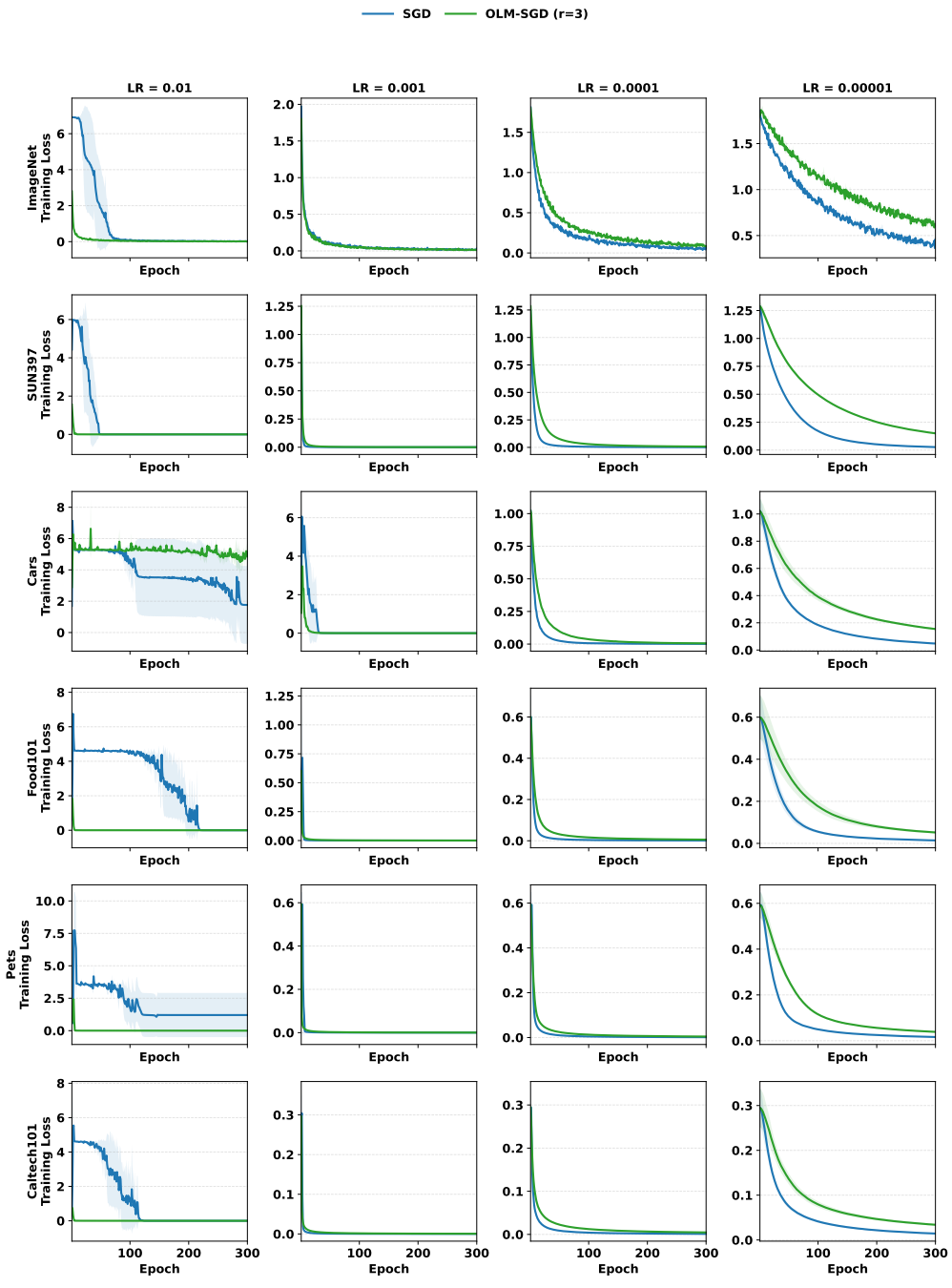

Figure 17: Training loss for the SGD and OLM-SGD trajectories in Figure 13. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines show means and shaded bands show one population standard deviation across the three seeds. ImageNet retains training augmentation under the shared protocol.

compared with eight for SGD. On Stanford Cars, the drop decreases from 1.93 points to 0.72. On Aircraft, it decreases from 2.24 to 1.56. ImageNet is still substantially less stable than the other targets. OLM-SGD loses 7.32 points there, compared with 13.36 for SGD. Excluding ImageNet, the ten-dataset mean drops are 0.31 and 0.65 points, respectively. Thus, ImageNet accounts for much of the aggregate degradation, but not the entire retention advantage.

At 10^{-5}, every OLM-SGD dataset-level mean drop is at most 0.43 points. Its overall drop of 0.11 points accompanies 71.75% final accuracy. Its small drop therefore accompanies useful adaptation. Nevertheless, the previously noted SGD wins on six datasets show why retention and absolute accuracy must be considered together.

A small drop can conceal an already failed run. At 10^{-2}, Adam's ImageNet drop is only 0.03 points, but its final accuracy is 0.11%. The metric is small because the best recorded checkpoint is already poor. Conversely, OLM-SGD loses 24.69 points on ImageNet at this rate despite having a much smaller dataset-average drop. The metric neither captures all damage relative to zero-shot CLIP nor guarantees useful adaptation. Together, the four analyses support a stronger accuracy–retention trade-off for OLM-SGD over the evaluated settings. They also establish SGD as an essential full-fine-tuning control. One of the main benefits is a wider useful range of learning rates and stopping times.

Table 39: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.01 (10^{-2}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	Adam	SGD	OLM-SGD ($r = 3$)
ImageNet	0.03 ± 0.02	2.91 ± 2.42	24.69 ± 0.49
Caltech101	92.83 ± 0.07	76.09 ± 10.44	0.07 ± 0.05
Oxford Pets	85.57 ± 0.22	82.65 ± 2.92	9.29 ± 6.93
Stanford Cars	0.28 ± 0.08	0.29 ± 0.28	2.83 ± 3.57
Flowers102	66.11 ± 0.48	60.52 ± 7.07	0.04 ± 0.06
Food101	84.24 ± 0.03	82.30 ± 0.20	0.00 ± 0.00
Aircraft	22.73 ± 0.07	21.19 ± 1.04	0.57 ± 0.39
SUN397	0.19 ± 0.01	0.29 ± 0.34	0.17 ± 0.17
DTD	41.94 ± 0.29	34.75 ± 5.73	0.12 ± 0.08
EuroSAT	29.93 ± 5.10	16.09 ± 7.40	0.17 ± 0.13
UCF101	64.10 ± 0.23	39.20 ± 8.75	0.03 ± 0.04
Average	44.36 ± 0.47	37.84 ± 0.77	3.45 ± 0.80

H.2 Comparison against SGD With Our Class-Unique Mini-Batching

This comparison asks whether class-unique minibatching alone explains the gains of OLM-SGD. SGD with class-unique minibatching uses the same batch construction as OLM-SGD but retains the ordinary mean-gradient update. It introduces neither per-example rank truncation nor polar reconstruction. We denote randomized SGD by SGD-R and class-unique SGD by SGD-CU in the tables. The common-horizon experiments compare all three methods at the same learning rate and training duration. All 396 trajectories contain complete test-accuracy and finite training-loss histories through epoch 300.

H.2.1 Cross-Dataset Transfer of the Selected Training Epoch

Higher accuracy with less dependence on the source duration. Figure 18 and Table 43 use 1-shot CLIP (ViT-B/16) at 10^{-4}. OLM-SGD attains 71.73% mean accuracy across the 110 off-diagonal transfers. SGD-CU and SGD-R attain 71.01% and 70.64%, respectively. The mean gap to the target's own best seed-mean checkpoint is 0.91 points for OLM-SGD. It is 1.11 for SGD-CU and 1.50 for SGD-R. OLM-SGD remains within one point of the target oracle on $85/110$ pairs, compared with $80/110$ and $75/110$ for the two SGD controls.

Table 40: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.001 (10^{-3}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	Adam	SGD	OLM-SGD ($r = 3$)
ImageNet	0.06 ± 0.01	22.99 ± 0.31	19.37 ± 0.63
Caltech101	92.67 ± 0.27	0.18 ± 0.07	0.20 ± 0.09
Oxford Pets	85.18 ± 0.35	0.14 ± 0.19	0.45 ± 0.15
Stanford Cars	0.33 ± 0.09	14.59 ± 8.09	0.03 ± 0.02
Flowers102	65.77 ± 0.57	10.43 ± 14.53	0.05 ± 0.02
Food101	84.24 ± 0.03	2.48 ± 1.95	0.05 ± 0.02
Aircraft	22.95 ± 0.09	18.78 ± 1.53	2.20 ± 0.39
SUN397	0.16 ± 0.06	0.98 ± 0.55	0.74 ± 0.18
DTD	42.32 ± 0.21	0.28 ± 0.12	0.14 ± 0.20
EuroSAT	27.56 ± 5.99	0.04 ± 0.00	0.84 ± 0.18
UCF101	64.21 ± 0.28	0.48 ± 0.25	0.29 ± 0.23
Average	44.13 ± 0.51	6.49 ± 1.35	2.22 ± 0.07

Table 41: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.0001 (10^{-4}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	Adam	SGD	OLM-SGD ($r = 3$)
ImageNet	13.10 ± 9.15	13.36 ± 0.81	7.32 ± 0.34
Caltech101	45.91 ± 12.92	0.42 ± 0.18	0.16 ± 0.10
Oxford Pets	81.54 ± 3.20	0.23 ± 0.16	0.07 ± 0.05
Stanford Cars	0.33 ± 0.18	1.93 ± 0.63	0.72 ± 0.13
Flowers102	54.38 ± 1.80	0.08 ± 0.07	0.04 ± 0.00
Food101	83.03 ± 0.31	0.26 ± 0.21	0.02 ± 0.02
Aircraft	22.52 ± 0.13	2.24 ± 0.33	1.56 ± 0.22
SUN397	4.09 ± 1.77	0.80 ± 0.01	0.28 ± 0.04
DTD	25.65 ± 1.40	0.24 ± 0.17	0.20 ± 0.17
EuroSAT	13.20 ± 1.58	0.21 ± 0.23	0.01 ± 0.02
UCF101	38.66 ± 3.23	0.11 ± 0.08	0.04 ± 0.04
Average	34.76 ± 1.87	1.81 ± 0.12	0.95 ± 0.06

The gain over SGD-CU combines a higher attainable accuracy with better retention under a transferred duration. The target-oracle average increases by 0.51 points, from 72.13% to 72.64%. The mean transfer penalty decreases by another 0.20 points. Using the unrounded values, these changes yield a 0.72-point transfer advantage. OLM-SGD also has a smaller worst-case oracle gap: 7.17 points, compared with 12.86 for SGD-CU and 26.15 for SGD-R.

The benefit extends beyond one target. Averaging over the ten source datasets for each target, OLM-SGD exceeds SGD-CU on ten of the eleven targets. The largest gains occur on EuroSAT (2.10 points), Flowers102 (1.60), and Stanford Cars (1.19). Caltech101 is close, with SGD-CU ahead by 0.11 points. OLM-SGD's source-selected epochs range from 18 to 292, whereas SGD-CU's range from 7 to 290. Hence, successful transfer does not require similar task-specific peak epochs. It requires the target trajectory to retain high accuracy across the selected durations. As defined in the shared protocol, this is a test-guided diagnostic of duration sensitivity rather than validation-selected early stopping.

Table 42: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.00001 (10^{-5}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean ± population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	Adam	SGD	OLM-SGD ($r = 3$)
ImageNet	36.85 ± 1.11	2.76 ± 0.27	0.43 ± 0.06
Caltech101	3.15 ± 0.74	0.07 ± 0.07	0.07 ± 0.04
Oxford Pets	10.55 ± 1.26	0.11 ± 0.14	0.02 ± 0.03
Stanford Cars	1.12 ± 1.47	0.69 ± 0.17	0.19 ± 0.16
Flowers102	0.88 ± 1.24	0.03 ± 0.04	0.01 ± 0.02
Food101	11.73 ± 0.70	0.22 ± 0.24	0.06 ± 0.09
Aircraft	3.75 ± 0.66	1.70 ± 0.04	0.29 ± 0.21
SUN397	0.07 ± 0.05	0.38 ± 0.10	0.02 ± 0.00
DTD	8.90 ± 2.33	0.12 ± 0.10	0.08 ± 0.06
EuroSAT	3.23 ± 2.27	0.06 ± 0.08	0.00 ± 0.01
UCF101	2.49 ± 0.34	0.00 ± 0.00	0.01 ± 0.01
Average	7.52 ± 0.86	0.56 ± 0.05	0.11 ± 0.00

Table 43: Epoch-transfer diagnostic for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Target oracle averages each dataset's best seed-mean accuracy. Transfer accuracy and oracle gaps average the 110 off-diagonal source–target pairs. Within 1 pp counts transfers no more than one point below the target oracle. The last column gives the largest oracle gap. All quantities are test-guided diagnostics.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp	Worst gap (pp)
SGD-R	72.14	70.64	1.50	75/110	26.15
SGD-CU	72.13	71.01	1.11	80/110	12.86
OLM-SGD	**72.64**	**71.73**	**0.91**	**85/110**	**7.17**

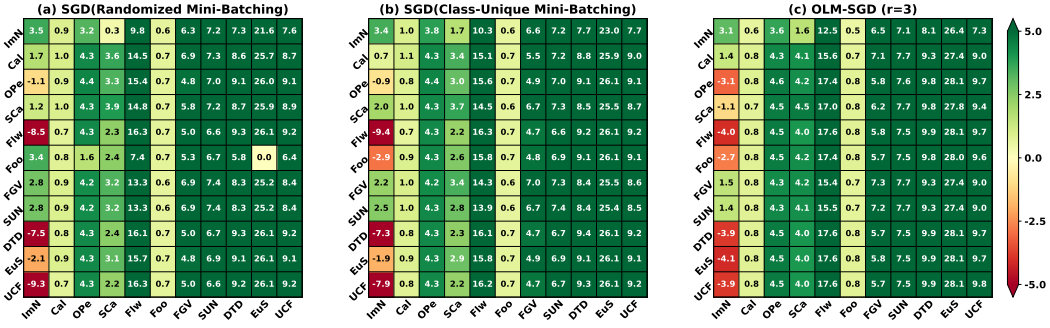

Figure 18: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Each source row selects the earliest epoch in 1–300 that maximizes mean *test* accuracy across seeds 1, 2, and 3. Each target column evaluates its independently trained trajectory at that epoch. Entries give the target accuracy change relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Colors saturate at ±5 points; annotations retain the rounded numerical values. OLM-SGD uses $r = 3$. Only the duration is transferred. Diagonal entries use target-specific oracle selection and are excluded from Table 43.

H.2.2 Long-Horizon Dynamics and Learning-Rate Sensitivity

Class-unique batching does not reproduce the stability gain. Figures 19 and 20 show the test-accuracy trajectories. At one shot, every support class has only one example. Randomized mini-batches are therefore already class-unique. The two SGD samplers have the same batching distri-

bution and update count per epoch, although their realized random orders can differ. Their close performance at the smaller rates is consistent with this equivalence. At 10^{-4}, SGD-R and SGD-CU reach 70.46% and 70.44% mean final accuracy. At 10^{-5}, both round to 71.46%. OLM-SGD reaches 71.75% at each rate. Thus, its advantage in this setting cannot be explained by removing repeated classes from a minibatch.

A wider useful learning-rate range. Table 44 summarizes all four rates. At 10^{-3}, OLM-SGD reaches 69.23% mean final accuracy, compared with 59.68% for SGD-CU and 58.67% for SGD-R. It exceeds both controls on all 11 datasets. Against SGD-CU, large gains occur on Stanford Cars (53.31% versus 24.07%), Aircraft (29.36% versus 5.01%), EuroSAT (70.71% versus 49.91%), and Flowers102 (86.26% versus 67.24%). At 10^{-4}, OLM-SGD again leads on every dataset. Its mean final advantage over SGD-CU is 1.31 points.

Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of mean final accuracy is 2.52 points for OLM-SGD. The corresponding ranges are 11.78 for SGD-CU and 12.79 for SGD-R. Including 10^{-2} gives ranges of 24.37, 58.81, and 59.99 points. At that largest rate, OLM-SGD still has a higher final average (47.38% versus 12.65% for SGD-CU), but all methods lose substantial accuracy relative to their smaller-rate results. The evidence supports a wider useful range, rather than insensitivity to arbitrarily large steps.

Table 44: Long-horizon summary for 1-shot CLIP (ViT-B/16). Final accuracy is evaluated at epoch 300. Best-to-final drops are computed within each seed before aggregation. For both quantities, datasets are averaged within each seed before reporting the mean and population standard deviation across seeds. The last column gives the mean accuracy change from epoch 30 to epoch 300. All 396 complete trajectories are included.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	SGD-R	11.47 ± 0.88	37.84 ± 0.77	$+7.60$
	SGD-CU	12.65 ± 1.68	37.11 ± 1.88	$+8.51$
	OLM-SGD	47.38 ± 0.43	3.45 ± 0.80	-0.18
10^{-3}	SGD-R	58.67 ± 2.09	6.49 ± 1.35	$+0.60$
	SGD-CU	59.68 ± 1.86	6.68 ± 1.81	-0.23
	OLM-SGD	69.23 ± 0.54	2.22 ± 0.07	-0.91
10^{-4}	SGD-R	70.46 ± 0.23	1.81 ± 0.12	-1.03
	SGD-CU	70.44 ± 0.23	1.82 ± 0.13	-1.04
	OLM-SGD	71.75 ± 0.26	0.95 ± 0.06	$+0.14$
10^{-5}	SGD-R	71.46 ± 0.02	0.56 ± 0.05	$+1.49$
	SGD-CU	71.46 ± 0.03	0.56 ± 0.03	$+1.51$
	OLM-SGD	71.75 ± 0.26	0.11 ± 0.00	$+3.85$

Retaining accuracy during prolonged training. At 10^{-4}, OLM-SGD improves from 71.61% at epoch 30 to 71.75% at epoch 300. SGD-CU instead decreases from 71.48% to 70.44%. Thus, the final gap is not merely inherited from the early checkpoint. It widens from 0.13 to 1.31 points as training continues. At 10^{-5}, adaptation is slower. Continuing from epoch 30 to epoch 300 adds 3.85 points for OLM-SGD and 1.51 for SGD-CU. SGD-CU remains competitive at this smallest rate and leads on six individual datasets. OLM-SGD's broader advantage is that strong average accuracy persists at the two larger rates without requiring such conservative steps.

Support fitting and generalization are distinct. Figures 21 and 22 show the training losses. At 10^{-4}, SGD-CU has lower final training loss than OLM-SGD on every dataset. Yet OLM-SGD has higher final test accuracy on every dataset. On Aircraft, SGD-CU reaches a final loss of approximately 1.33×10^{-3} and 28.58% accuracy. OLM-SGD retains 29.70% accuracy with a larger loss of 6.85×10^{-3}. At 10^{-3}, the distinction is stronger. Both methods nearly fit the support set on Aircraft, with losses below 5×10^{-4}, but their final accuracies are 5.01% and 29.36%. Thus, the spectral merge changes generalization behavior, not simply whether the support labels can be fitted.

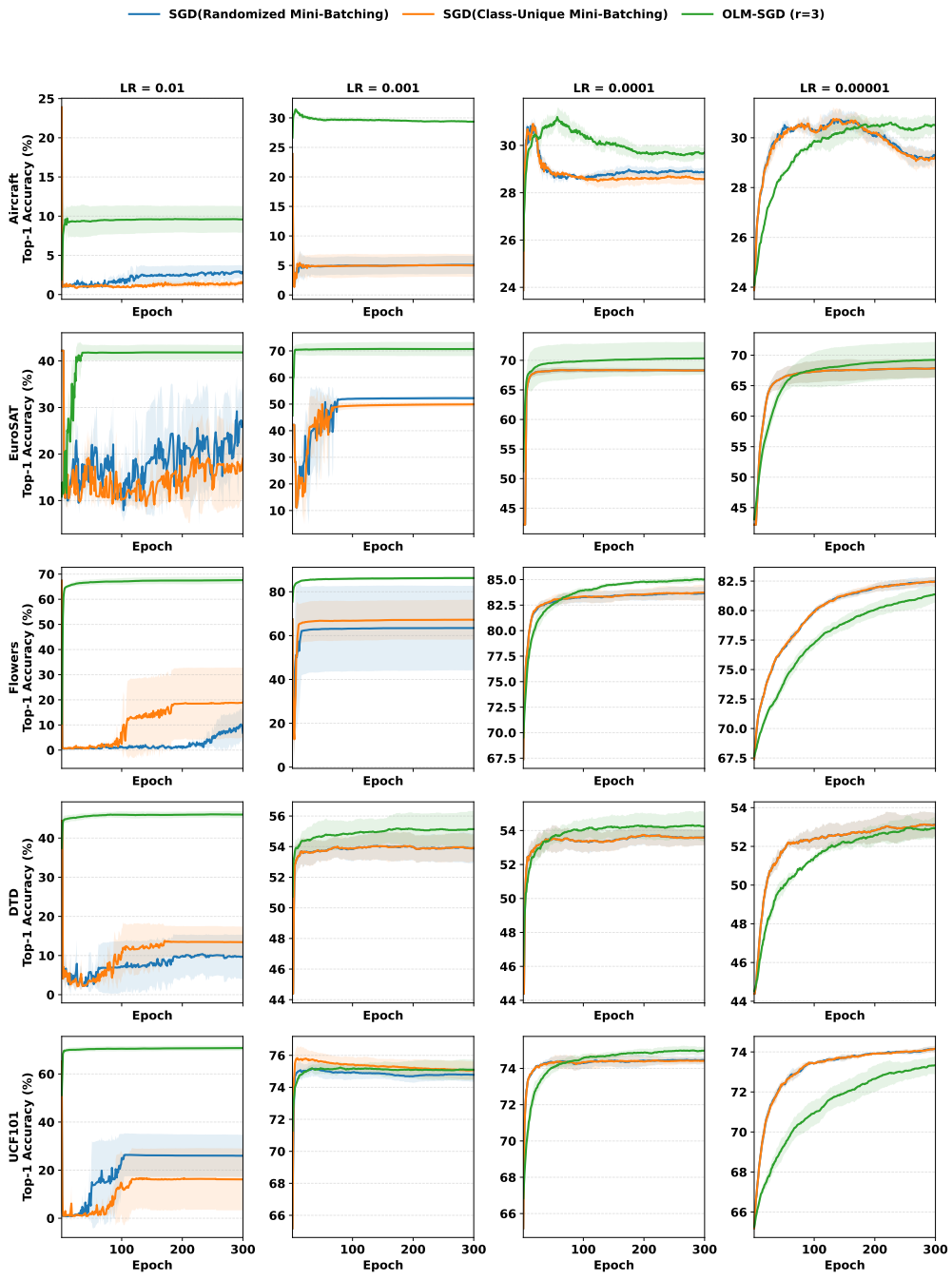

Figure 19: Test-accuracy trajectories for SGD with randomized minibatching, SGD with class-unique minibatching, and OLM-SGD ($r = 3$). Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The shared protocol uses 1-shot CLIP (ViT-B/16) over 300 epochs. Lines show means and shaded bands show one population standard deviation across seeds 1, 2, and 3.

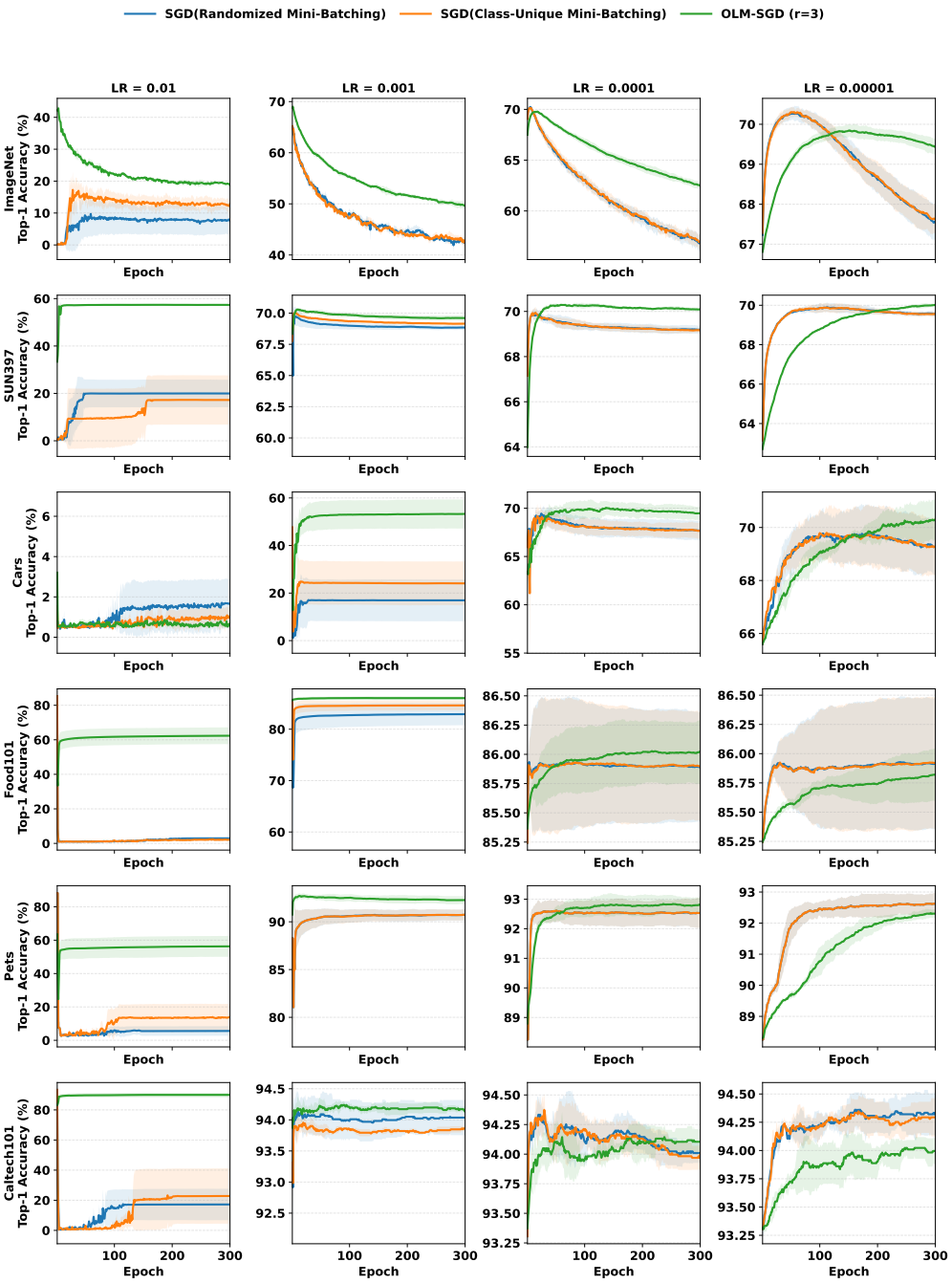

Figure 20: Test-accuracy trajectories continuing Figure 19. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. ImageNet retains training augmentation under the shared protocol.

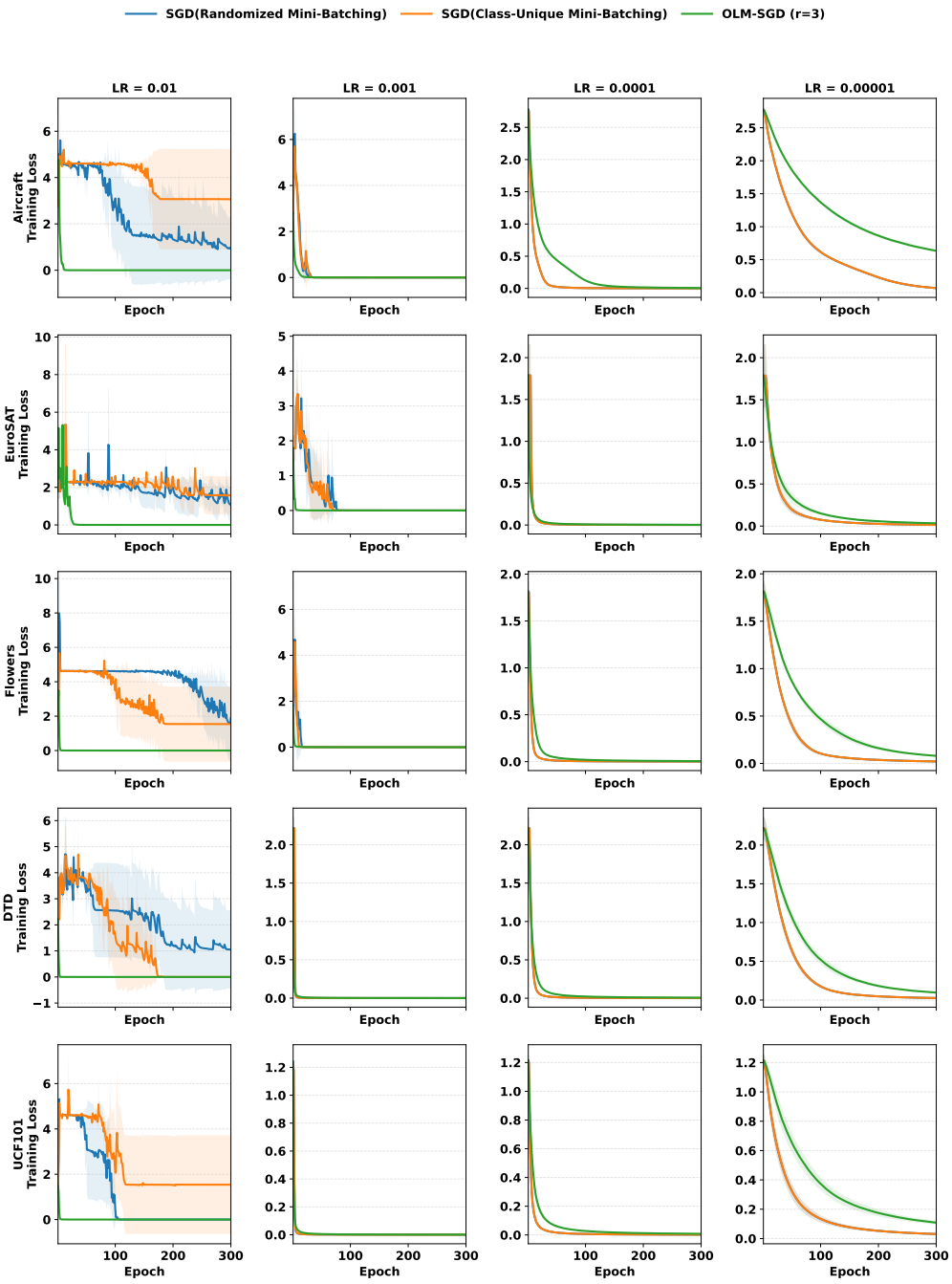

Figure 21: Training loss for the three methods in Figure 19. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across the three seeds. These are support-set training losses, not test losses. All displayed loss histories are finite and complete.

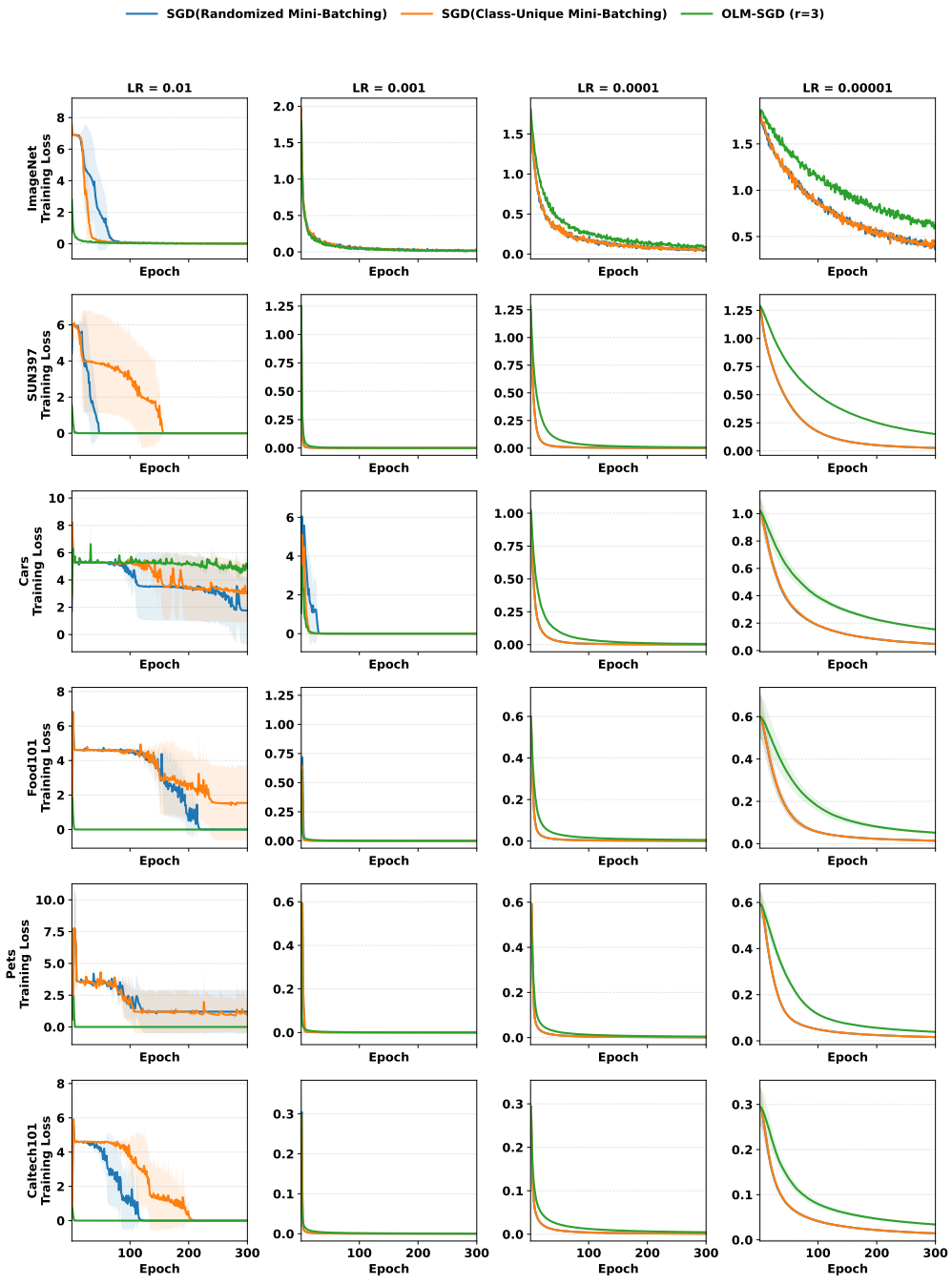

Figure 22: Training loss for the three methods in Figure 20. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. ImageNet retains training augmentation. All displayed loss histories are finite and complete.

H.2.3 BEST-TO-FINAL TEST-ACCURACY DROP

Smaller mean degradation at every tested rate. Tables 45–48 quantify accuracy retention. Drops are computed within each seed before aggregation, as specified in the shared protocol. OLM-SGD has the smallest dataset-average drop at all four learning rates. Its drops are 3.45, 2.22, 0.95, and 0.11 points at 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. SGD-CU loses 37.11, 6.68, 1.82, and 0.56 points. SGD-R loses 37.84, 6.49, 1.81, and 0.56 points. The two SGD controls exhibit similar aggregate degradation, whereas OLM-SGD improves retention throughout the sweep.

The advantage is not confined to the largest outlier. At 10^{-4}, OLM-SGD has a smaller mean drop than SGD-CU on all 11 datasets. The aggregate reduction is approximately 48%. On Stanford Cars, the drop decreases from 1.86 to 0.72 points. On Aircraft, it decreases from 2.51 to 1.56. ImageNet has the largest absolute drop for both methods, but OLM-SGD reduces it from 13.24 to 7.32 points. Excluding ImageNet, the ten-dataset averages are still 0.68 points for SGD-CU and 0.31 for OLM-SGD. Thus, improved retention is shared across targets rather than driven only by ImageNet.

At 10^{-3}, the Stanford Cars drop decreases from 23.59 ± 15.78 points with SGD-CU to 0.03 ± 0.02 with OLM-SGD. Aircraft decreases from 18.87 ± 1.91 to 2.20 ± 0.39 points. These reductions accompany the higher final accuracies reported above. At 10^{-5}, OLM-SGD's mean drop is at most 0.43 points on every dataset. Its overall drop of 0.11 points accompanies 71.75% final accuracy. The retention improvement remains visible when the SGD controls are already competitive in absolute accuracy.

Implication for the role of the merge. The matched-rate experiments provide the central evidence. At one shot, class uniqueness is already satisfied by ordinary randomized minibatches. Nevertheless, OLM-SGD improves final accuracy on every dataset at 10^{-3} and 10^{-4}. It also reduces the average transfer penalty and the mean best-to-final drop. Class-unique batch construction alone is insufficient to explain these gains. The results support structured gradient merging as a central component of OLM-SGD's robustness to learning rate and training duration.

Table 45: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.01 (10^{-2}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	SGD (Randomized Mini-Batching)	SGD (Class-Unique Mini-Batching)	OLM-SGD ($r = 3$)
ImageNet	2.91 ± 2.42	5.56 ± 2.66	24.69 ± 0.49
Caltech101	76.09 ± 10.44	70.48 ± 18.40	0.07 ± 0.05
Oxford Pets	82.65 ± 2.92	74.37 ± 7.83	9.29 ± 6.93
Stanford Cars	0.29 ± 0.28	0.58 ± 0.36	2.83 ± 3.57
Flowers102	60.52 ± 7.07	48.36 ± 13.73	0.04 ± 0.06
Food101	82.30 ± 0.20	83.09 ± 0.71	0.00 ± 0.00
Aircraft	21.19 ± 1.04	22.31 ± 0.32	0.57 ± 0.39
SUN397	0.29 ± 0.34	0.26 ± 0.29	0.17 ± 0.17
DTD	34.75 ± 5.73	30.97 ± 3.98	0.12 ± 0.08
EuroSAT	16.09 ± 7.40	23.21 ± 8.58	0.17 ± 0.13
UCF101	39.20 ± 8.75	49.03 ± 12.88	0.03 ± 0.04
Average	37.84 ± 0.77	37.11 ± 1.88	3.45 ± 0.80

Table 46: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.001 (10^{-3}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	SGD (Randomized Mini-Batching)	SGD (Class-Unique Mini-Batching)	OLM-SGD ($r = 3$)
ImageNet	22.99 ± 0.31	22.57 ± 0.14	19.37 ± 0.63
Caltech101	0.18 ± 0.07	0.19 ± 0.12	0.20 ± 0.09
Oxford Pets	0.14 ± 0.19	0.11 ± 0.15	0.45 ± 0.15
Stanford Cars	14.59 ± 8.09	23.59 ± 15.78	0.03 ± 0.02
Flowers102	10.43 ± 14.53	4.40 ± 6.02	0.05 ± 0.02
Food101	2.48 ± 1.95	0.75 ± 0.85	0.05 ± 0.02
Aircraft	18.78 ± 1.53	18.87 ± 1.91	2.20 ± 0.39
SUN397	0.98 ± 0.55	1.00 ± 0.27	0.74 ± 0.18
DTD	0.28 ± 0.12	0.24 ± 0.10	0.14 ± 0.20
EuroSAT	0.04 ± 0.00	0.92 ± 0.75	0.84 ± 0.18
UCF101	0.48 ± 0.25	0.87 ± 0.36	0.29 ± 0.23
Average	6.49 ± 1.35	6.68 ± 1.81	2.22 ± 0.07

Table 47: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.0001 (10^{-4}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	SGD (Randomized Mini-Batching)	SGD (Class-Unique Mini-Batching)	OLM-SGD ($r = 3$)
ImageNet	13.36 ± 0.81	13.24 ± 0.42	7.32 ± 0.34
Caltech101	0.42 ± 0.18	0.39 ± 0.17	0.16 ± 0.10
Oxford Pets	0.23 ± 0.16	0.22 ± 0.17	0.07 ± 0.05
Stanford Cars	1.93 ± 0.63	1.86 ± 0.83	0.72 ± 0.13
Flowers102	0.08 ± 0.07	0.11 ± 0.13	0.04 ± 0.00
Food101	0.26 ± 0.21	0.28 ± 0.24	0.02 ± 0.02
Aircraft	2.24 ± 0.33	2.51 ± 0.12	1.56 ± 0.22
SUN397	0.80 ± 0.01	0.82 ± 0.11	0.28 ± 0.04
DTD	0.24 ± 0.17	0.28 ± 0.14	0.20 ± 0.17
EuroSAT	0.21 ± 0.23	0.21 ± 0.23	0.01 ± 0.02
UCF101	0.11 ± 0.08	0.14 ± 0.09	0.04 ± 0.04
Average	1.81 ± 0.12	1.82 ± 0.13	0.95 ± 0.06

Table 48: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.00001 (10^{-5}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	SGD (Randomized Mini-Batching)	SGD (Class-Unique Mini-Batching)	OLM-SGD ($r = 3$)
ImageNet	2.76 ± 0.27	2.71 ± 0.23	0.43 ± 0.06
Caltech101	0.07 ± 0.07	0.09 ± 0.05	0.07 ± 0.04
Oxford Pets	0.11 ± 0.14	0.11 ± 0.14	0.02 ± 0.03
Stanford Cars	0.69 ± 0.17	0.66 ± 0.14	0.19 ± 0.16
Flowers102	0.03 ± 0.04	0.01 ± 0.02	0.01 ± 0.02
Food101	0.22 ± 0.24	0.21 ± 0.25	0.06 ± 0.09
Aircraft	1.70 ± 0.04	1.75 ± 0.11	0.29 ± 0.21
SUN397	0.38 ± 0.10	0.40 ± 0.10	0.02 ± 0.00
DTD	0.12 ± 0.10	0.12 ± 0.10	0.08 ± 0.06
EuroSAT	0.06 ± 0.08	0.07 ± 0.08	0.00 ± 0.01
UCF101	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.01
Average	0.56 ± 0.05	0.56 ± 0.03	0.11 ± 0.00

H.3 Comparison against CTF-SGD

CTF-SGD tests whether controlling the magnitude of the ordinary SGD update is sufficient to recover the accuracy and stability of OLM-SGD. As defined in Appendix C, it estimates the cross-term-free gradient energy from two microbatches. A single global coefficient $\alpha_t^{\mathrm{CTF}} \in [0, 1]$ then rescales the minibatch mean gradient. The update remains collinear with SGD and cannot amplify its magnitude. Unlike OLM-SGD, CTF-SGD applies neither individual-gradient rank truncation nor polar reconstruction. It is a scale-control baseline, not an exact match to the OLM-SGD update norm. All 396 trajectories for SGD, CTF-SGD, and OLM-SGD contain complete test-accuracy and finite training-loss histories through epoch 300.

H.3.1 Few-Shot Accuracy across Backbones and Shot Levels

Average accuracy under the reported schedules. Tables 50–52 report the three-backbone results. Table 49 summarizes their Average columns. CTF-SGD uses randomized minibatches and runs for 30 epochs at 10^{-3}. SGD runs for 30 epochs at 10^{-4}, while OLM-SGD runs for 70 epochs at 10^{-4}. These tables measure adaptation under the reported schedules. The subsequent trajectory experiments match the learning rate and training horizon.

OLM-SGD exceeds CTF-SGD at every shot level on ViT-B/32 and ViT-B/16. On ViT-B/32, the gains at 1, 2, 4, 8, and 16 shots are 5.7, 3.9, 2.7, 1.7, and 1.0 points. On ViT-B/16, they are 6.1, 4.6, 2.8, 1.9, and 1.0 points. Averaging these displayed margins gives 3.00 and 3.28 points, respectively. The largest improvements occur with the fewest labels. The positive margins at 16 shots show that the benefit persists as the support set grows.

Table 49: Dataset-average few-shot accuracy (%) from Tables 50–52. We reproduce the displayed Average means. The final column subtracts CTF-SGD from OLM-SGD, in percentage points. Bold marks the highest displayed average in each row.

Backbone	Shots	SGD	CTF-SGD	OLM-SGD	OLM-SGD − CTF-SGD
ViT-B/32	1	66.9	61.5	**67.2**	+5.7
	2	69.8	66.5	**70.4**	+3.9
	4	72.3	70.4	**73.1**	+2.7
	8	74.9	74.1	**75.8**	+1.7
	16	77.6	77.5	**78.5**	+1.0
ViT-B/16	1	71.5	65.9	**72.0**	+6.1
	2	74.5	70.5	**75.1**	+4.6
	4	76.8	75.0	**77.8**	+2.8
	8	79.1	78.4	**80.3**	+1.9
	16	81.6	81.6	**82.6**	+1.0
ViT-L/14	1	**77.8**	77.7	77.4	−0.3
	2	80.0	**80.4**	80.1	−0.3
	4	82.1	**82.5**	**82.5**	0.0
	8	84.4	**84.9**	84.7	−0.2
	16	86.3	86.7	**86.8**	+0.1

Dataset-level gains and backbone dependence. Using the displayed means, OLM-SGD exceeds CTF-SGD in $42/55$ dataset–shot settings on ViT-B/32 and $43/55$ on ViT-B/16. There are five and one additional ties, respectively. Across all three backbones, OLM-SGD leads on ImageNet and Stanford Cars in all 15 settings. It also leads on Aircraft in $13/15$ settings. For 1-shot ViT-B/32, Stanford Cars improves from 36.7% to 62.6%. For 1-shot ViT-B/16, Aircraft improves from 17.2% to 30.7%. At 16 shots on the same backbone, the Aircraft gain remains 4.9 points, from 48.0% to 52.9%.

On ViT-L/14, the dataset averages are much closer. CTF-SGD leads by 0.3, 0.3, and 0.2 points at 1, 2, and 8 shots. The methods tie at 4 shots, and OLM-SGD leads at 16 shots with 86.8% versus 86.7%. Thus, the fixed-schedule gains are strongest on the two smaller backbones. The matched 1-shot ViT-B/16 trajectories below establish whether the advantage also holds when learning rate and duration are controlled.

Table 50: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average is the equally weighted mean of the 11 dataset means. SGD uses 30 epochs at 10^{-4}, CTF-SGD uses 30 epochs at 10^{-3}, and OLM-SGD uses 70 epochs at 10^{-4}.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	SGD	61.2 ± 0.3	67.0 ± 0.2	22.3 ± 0.5	62.6 ± 5.5	61.4 ± 0.6	80.3 ± 0.2	89.0 ± 0.6	77.7 ± 0.8	93.6 ± 0.6	49.7 ± 0.3	70.5 ± 0.5	66.9
	CTF-SGD	46.2 ± 0.3	66.7 ± 0.2	12.2 ± 0.8	60.9 ± 7.8	36.7 ± 3.9	77.1 ± 2.7	87.4 ± 1.3	75.3 ± 2.3	93.4 ± 0.8	49.2 ± 0.9	71.0 ± 0.9	61.5
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.7	50.7 ± 1.9	70.2 ± 0.4	67.2
2	SGD	61.6 ± 0.2	68.2 ± 0.2	23.9 ± 0.2	75.3 ± 2.2	63.3 ± 0.1	80.6 ± 0.1	89.8 ± 0.5	82.6 ± 0.8	94.5 ± 0.1	54.6 ± 1.6	72.9 ± 0.4	69.8
	CTF-SGD	48.9 ± 1.3	68.4 ± 0.4	14.5 ± 0.3	76.8 ± 2.8	50.9 ± 1.8	79.3 ± 0.3	88.9 ± 0.6	81.6 ± 1.2	94.0 ± 0.2	54.1 ± 1.8	73.8 ± 0.3	66.5
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.8 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.3 ± 0.9	70.4
4	SGD	62.9 ± 0.3	69.6 ± 0.3	27.4 ± 0.3	81.2 ± 1.3	66.3 ± 0.5	80.8 ± 0.2	90.6 ± 0.4	87.2 ± 0.8	94.8 ± 0.1	58.4 ± 0.5	75.6 ± 0.7	72.3
	CTF-SGD	51.6 ± 0.8	70.0 ± 0.2	19.2 ± 1.2	84.5 ± 3.1	61.2 ± 1.1	79.5 ± 0.6	89.6 ± 0.6	87.9 ± 0.6	94.8 ± 0.1	59.6 ± 0.4	76.9 ± 1.0	70.4
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.2 ± 0.6	81.3 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1
8	SGD	64.0 ± 0.2	71.0 ± 0.2	32.4 ± 0.7	88.5 ± 0.7	70.1 ± 0.3	81.0 ± 0.2	90.9 ± 0.4	90.5 ± 0.6	95.2 ± 0.4	62.4 ± 0.4	77.8 ± 0.1	74.9
	CTF-SGD	54.5 ± 0.3	72.0 ± 0.2	28.9 ± 0.9	90.0 ± 0.9	69.2 ± 0.3	79.9 ± 0.1	90.4 ± 0.4	91.4 ± 0.5	95.1 ± 0.5	64.4 ± 0.3	79.3 ± 0.2	74.1
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.7 ± 0.1	91.2 ± 0.5	92.0 ± 0.1	95.2 ± 0.2	64.6 ± 0.7	78.5 ± 0.4	75.8
16	SGD	66.6 ± 0.2	72.3 ± 0.1	39.5 ± 0.3	91.9 ± 0.4	74.5 ± 0.3	81.3 ± 0.0	90.9 ± 0.5	93.8 ± 0.1	95.5 ± 0.1	67.1 ± 0.6	79.9 ± 0.1	77.6
	CTF-SGD	58.5 ± 0.1	73.5 ± 0.0	40.1 ± 0.3	93.4 ± 0.4	74.9 ± 0.2	80.4 ± 0.3	90.2 ± 0.7	95.0 ± 0.6	95.6 ± 0.1	68.8 ± 0.3	81.8 ± 0.3	77.5
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5

Table 51: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average is the equally weighted mean of the 11 dataset means. SGD uses 30 epochs at 10^{-4}, CTF-SGD uses 30 epochs at 10^{-3}, and OLM-SGD uses 70 epochs at 10^{-4}.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	SGD	67.5 ± 0.2	69.6 ± 0.2	29.4 ± 0.2	68.1 ± 0.6	69.3 ± 0.9	85.9 ± 0.6	92.6 ± 0.4	82.5 ± 0.4	94.2 ± 0.1	53.2 ± 0.7	74.2 ± 0.1	71.5
	CTF-SGD	54.6 ± 0.8	69.7 ± 0.0	17.2 ± 1.0	45.5 ± 16.8	56.6 ± 2.9	85.4 ± 0.3	90.7 ± 0.3	81.5 ± 1.1	94.2 ± 0.1	53.7 ± 0.6	75.6 ± 1.1	65.9
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.4	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0
2	SGD	67.9 ± 0.2	71.4 ± 0.1	32.3 ± 1.2	77.9 ± 1.1	71.7 ± 0.4	86.0 ± 0.1	92.8 ± 0.4	88.6 ± 0.6	94.5 ± 0.3	59.5 ± 1.0	77.3 ± 0.3	74.5
	CTF-SGD	56.2 ± 0.7	71.6 ± 0.1	19.1 ± 2.2	75.2 ± 4.5	60.1 ± 3.0	85.4 ± 0.1	90.5 ± 2.5	85.4 ± 2.2	94.8 ± 0.1	59.4 ± 1.7	78.1 ± 0.3	70.5
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.1 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1
4	SGD	68.9 ± 0.1	72.4 ± 0.2	36.3 ± 0.6	82.6 ± 1.5	74.7 ± 0.2	86.5 ± 0.1	93.4 ± 0.2	91.9 ± 0.5	95.3 ± 0.1	63.1 ± 0.4	79.6 ± 0.2	76.8
	CTF-SGD	59.2 ± 0.2	73.2 ± 0.0	27.4 ± 0.8	84.4 ± 3.0	70.1 ± 0.2	86.2 ± 0.1	93.4 ± 0.3	90.2 ± 1.1	95.3 ± 0.4	64.7 ± 0.7	80.7 ± 0.7	75.0
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.1	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8
8	SGD	70.1 ± 0.3	73.7 ± 0.3	42.2 ± 0.6	88.3 ± 0.7	77.7 ± 0.4	86.8 ± 0.3	93.5 ± 0.3	94.1 ± 0.2	95.6 ± 0.1	66.3 ± 0.8	81.5 ± 0.3	79.1
	CTF-SGD	62.1 ± 0.7	74.6 ± 0.2	37.5 ± 0.6	89.0 ± 3.2	77.6 ± 0.1	85.7 ± 0.2	93.3 ± 0.3	95.6 ± 0.3	96.1 ± 0.2	68.6 ± 0.5	82.8 ± 0.2	78.4
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.2	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3
16	SGD	72.1 ± 0.1	74.8 ± 0.1	48.6 ± 0.3	92.1 ± 0.8	81.3 ± 0.1	87.0 ± 0.1	94.0 ± 0.1	96.5 ± 0.1	96.2 ± 0.2	71.0 ± 0.4	83.7 ± 0.2	81.6
	CTF-SGD	65.8 ± 0.3	76.1 ± 0.3	48.0 ± 0.5	93.8 ± 0.6	82.5 ± 0.3	86.1 ± 0.3	93.5 ± 0.2	97.5 ± 0.1	96.3 ± 0.2	72.5 ± 0.9	85.8 ± 0.3	81.6
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.1	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6

Table 52: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. Few-shot entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average is the equally weighted mean of the 11 dataset means. SGD uses 30 epochs at 10^{-4}, CTF-SGD uses 30 epochs at 10^{-3}, and OLM-SGD uses 70 epochs at 10^{-4}.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	SGD	76.6 ± 0.1	73.9 ± 0.0	40.0 ± 1.4	73.5 ± 1.0	80.3 ± 0.3	90.9 ± 0.1	94.4 ± 0.2	88.7 ± 0.4	96.0 ± 0.6	60.1 ± 0.9	81.4 ± 0.6	77.8
	CTF-SGD	71.3 ± 0.5	74.7 ± 0.2	38.3 ± 0.8	74.4 ± 3.1	79.8 ± 0.1	90.9 ± 0.1	94.5 ± 0.1	90.9 ± 0.8	96.6 ± 0.4	61.0 ± 0.7	82.0 ± 0.7	77.7
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.5 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4
2	SGD	76.8 ± 0.3	75.8 ± 0.3	43.4 ± 1.7	79.4 ± 1.6	82.2 ± 0.3	91.1 ± 0.2	94.5 ± 0.0	92.9 ± 0.4	96.7 ± 0.2	65.3 ± 0.7	82.5 ± 0.8	80.0
	CTF-SGD	71.6 ± 0.3	76.4 ± 0.2	38.7 ± 2.7	87.2 ± 1.3	82.0 ± 0.0	91.2 ± 0.2	94.8 ± 0.1	94.6 ± 0.5	97.2 ± 0.1	67.4 ± 1.1	83.7 ± 0.8	80.4
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1
4	SGD	77.7 ± 0.1	77.0 ± 0.2	49.0 ± 1.3	83.1 ± 2.1	83.9 ± 0.2	91.4 ± 0.1	95.1 ± 0.2	95.6 ± 0.3	97.6 ± 0.4	68.2 ± 0.9	84.8 ± 0.4	82.1
	CTF-SGD	73.0 ± 0.2	77.7 ± 0.1	47.4 ± 1.3	86.9 ± 0.9	84.2 ± 0.2	91.5 ± 0.1	95.4 ± 0.2	97.4 ± 0.2	97.7 ± 0.4	70.0 ± 1.2	86.3 ± 0.6	82.5
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.2	91.3 ± 0.1	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.3	68.7 ± 0.9	84.8 ± 0.5	82.5
8	SGD	78.4 ± 0.2	78.5 ± 0.5	56.4 ± 0.6	89.0 ± 1.0	86.5 ± 0.3	91.7 ± 0.1	95.2 ± 0.2	97.0 ± 0.1	97.7 ± 0.3	72.0 ± 0.6	86.3 ± 0.2	84.4
	CTF-SGD	74.5 ± 0.0	79.2 ± 0.4	57.0 ± 1.2	91.9 ± 0.9	87.0 ± 0.2	91.8 ± 0.1	95.8 ± 0.1	98.0 ± 0.0	97.7 ± 0.3	73.4 ± 0.4	87.5 ± 0.4	84.9
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7
16	SGD	79.4 ± 0.1	79.6 ± 0.1	62.6 ± 0.2	91.8 ± 0.8	88.0 ± 0.3	91.8 ± 0.1	95.8 ± 0.1	98.3 ± 0.1	97.9 ± 0.1	75.7 ± 0.4	88.0 ± 0.2	86.3
	CTF-SGD	76.5 ± 0.1	80.2 ± 0.1	64.1 ± 0.6	94.1 ± 0.2	88.9 ± 0.2	91.9 ± 0.1	96.0 ± 0.1	98.8 ± 0.1	97.9 ± 0.1	77.0 ± 0.3	88.9 ± 0.3	86.7
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8

H.3.2 Cross-Dataset Transfer of the Selected Training Epoch

Higher transfer accuracy and better accuracy retention. Figure 23 and Table 53 evaluate the source-selected durations at 10^{-4}. OLM-SGD reaches 71.73% mean accuracy across the 110 off-diagonal transfers. CTF-SGD and SGD reach 70.74% and 70.64%, respectively. OLM-SGD's mean gap to the target oracle is 0.91 points, compared with 1.36 for CTF-SGD and 1.50 for SGD. It remains within one point of the oracle on $85/110$ pairs, compared with $78/110$ and $75/110$ for the two controls.

The 0.99-point transfer advantage over CTF-SGD combines two effects. OLM-SGD's target-oracle average is 0.53 points higher. The penalty for transferring the duration is another 0.46 points smaller. Hence, the gain is not explained only by a higher attainable peak. OLM-SGD also preserves more of that peak under source-selected stopping times. Its largest oracle gap is 7.17 points, compared with 14.02 for CTF-SGD. The largest CTF-SGD gap occurs when ImageNet selects epoch 6 for EuroSAT, which benefits from longer adaptation.

Table 53: Epoch-transfer diagnostic for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. The target oracle averages each dataset's best seed-mean accuracy. Transfer accuracy and oracle gaps average the 110 off-diagonal source–target pairs. The final column counts pairs within one percentage point of the target oracle. These are test-guided diagnostics.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp
SGD	72.14	70.64	1.50	75/110
CTF-SGD	72.11	70.74	1.36	78/110
OLM-SGD	**72.64**	**71.73**	**0.91**	**85/110**

The improvement extends across targets. Averaging the ten transferred durations for each target, OLM-SGD exceeds CTF-SGD on ten of the eleven datasets. The largest gains occur on EuroSAT (3.56 points), Flowers102 (1.91), Stanford Cars (1.27), and ImageNet (1.07). The remaining target, Caltech101, differs by only 0.09 points in favor of CTF-SGD. OLM-SGD's source-selected epochs range from 18 to 292, whereas CTF-SGD's range from 6 to 271. Successful transfer does not require similar peak epochs across datasets. It requires high target accuracy across the transferred durations. Following the shared protocol, these results diagnose duration sensitivity through test-guided epoch selection. They do not use validation-selected early stopping.

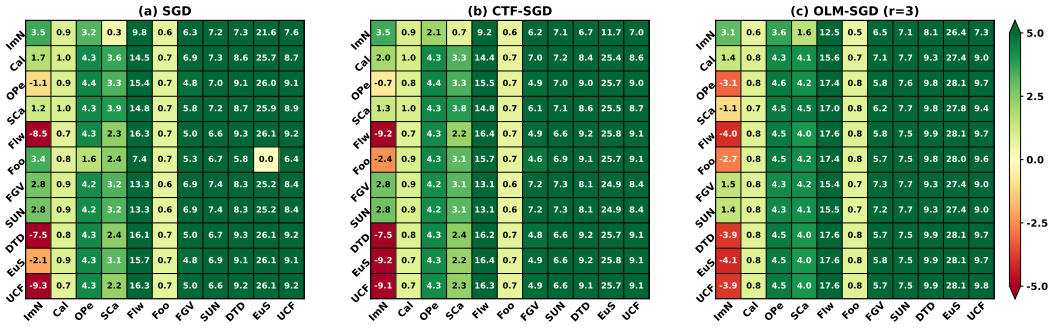

Figure 23: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Each source row selects the earliest epoch in 1–300 that maximizes mean *test* accuracy across seeds 1, 2, and 3. Each target column evaluates its independently trained trajectory at that epoch. Entries give the target accuracy change relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Colors saturate at ± 5 points, while annotations retain the rounded numerical values. OLM-SGD uses $r = 3$. Only the duration is transferred. Diagonal entries use target-specific oracle selection and are excluded from Table 53.

H.3.3 Long-Horizon Dynamics and Learning-Rate Sensitivity

Scalar shrinkage recovers only part of the robustness. Figures 24 and 25 show the test-accuracy trajectories. Table 54 summarizes their endpoints. At 10^{-3}, CTF-SGD improves mean final accuracy from SGD's 58.67% to 64.67%. OLM-SGD reaches 69.23%, adding another 4.56 points. It exceeds CTF-SGD on all 11 datasets. Large gains occur on EuroSAT (70.71% versus 56.47%), Stanford Cars (53.31% versus 42.65%), and Aircraft (29.36% versus 18.79%). At 10^{-4}, OLM-SGD again leads on every dataset, with mean final accuracy of 71.75% versus 70.42%.

At the stress-test rate of 10^{-2}, CTF-SGD reaches 21.06% final accuracy, compared with 11.47% for SGD. OLM-SGD retains 47.38%. Although this extreme rate reduces accuracy substantially, the relative advantage remains large. Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of mean final accuracy is 2.52 points for OLM-SGD. It is 6.76 for CTF-SGD and 12.79 for SGD. Including 10^{-2} expands these ranges to 24.37, 50.37, and 59.99 points. Thus, the reduced learning-rate sensitivity is visible both within the more effective range and across the full sweep.

Table 54: Long-horizon summary for 1-shot CLIP (ViT-B/16). Final accuracy is evaluated at epoch 300. Best-to-final drops are computed within each seed before aggregation. For both quantities, we first average the 11 datasets within each seed and then report mean $\pm$ population standard deviation across seeds. The last column gives the mean accuracy change from epoch 30 to epoch 300.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	SGD	11.47 ± 0.88	37.84 ± 0.77	$+7.60$
	CTF-SGD	21.06 ± 0.31	30.45 ± 0.33	$+9.62$
	OLM-SGD	47.38 ± 0.43	3.45 ± 0.80	-0.18
10^{-3}	SGD	58.67 ± 2.09	6.49 ± 1.35	$+0.60$
	CTF-SGD	64.67 ± 1.87	4.55 ± 2.23	-0.39
	OLM-SGD	69.23 ± 0.54	2.22 ± 0.07	-0.91
10^{-4}	SGD	70.46 ± 0.23	1.81 ± 0.12	-1.03
	CTF-SGD	70.42 ± 0.19	1.80 ± 0.14	-1.02
	OLM-SGD	71.75 ± 0.26	0.95 ± 0.06	$+0.14$
10^{-5}	SGD	71.46 ± 0.02	0.56 ± 0.05	$+1.49$
	CTF-SGD	71.43 ± 0.03	0.58 ± 0.06	$+1.57$
	OLM-SGD	71.75 ± 0.26	0.11 ± 0.00	$+3.85$

Retaining performance beyond the initial adaptation phase. At 10^{-4}, OLM-SGD's mean accuracy increases by 0.14 points between epochs 30 and 300. CTF-SGD instead loses 1.02 points. The gap between the methods grows from 0.17 points at epoch 30 to 1.33 points at epoch 300. Starting from epoch 70 gives the same retention ordering: OLM-SGD loses 0.27 points, whereas CTF-SGD loses 0.71. Thus, the benefit is not confined to the early adaptation phase.

At 10^{-5}, OLM-SGD adapts more slowly. Its epoch-30 average is 67.89%, compared with 69.86% for CTF-SGD. Continuing to epoch 300 adds 3.85 and 1.57 points, respectively. OLM-SGD finishes at 71.75%, compared with 71.43% for CTF-SGD. The smaller learning rate narrows the final-accuracy gap, but it requires a longer horizon to realize OLM-SGD's adaptation gains.

Lower support loss does not explain the accuracy advantage. Figures 26 and 27 separate support fitting from test performance. At 10^{-4}, CTF-SGD has lower mean final training loss than OLM-SGD on every dataset. Nevertheless, its final test accuracy is lower on every dataset. On Aircraft, the losses are approximately 1.38×10^{-3} and 6.85×10^{-3} for CTF-SGD and OLM-SGD, while test accuracies are 28.80% and 29.70%. On ImageNet, the losses are 0.0796 and 0.0844, whereas test accuracies are 56.93% and 62.48%. At 10^{-3}, CTF-SGD again has lower final loss on ten datasets but lower final accuracy on all eleven. These results show that stronger support fitting is not sufficient to reproduce the generalization of the spectral merge.

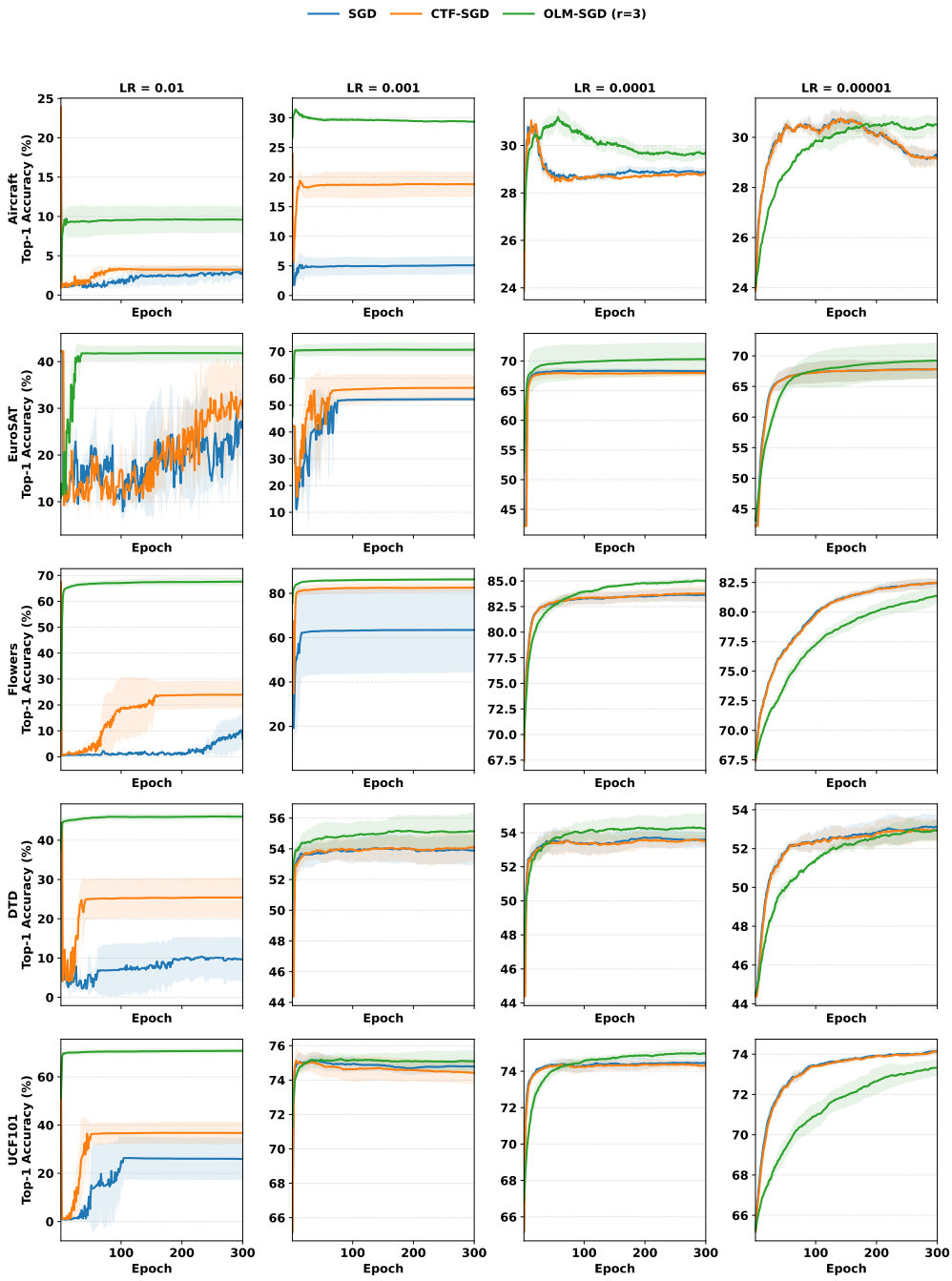

Figure 24: Test-accuracy trajectories for SGD, CTF-SGD, and OLM-SGD ($r = 3$). Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The shared protocol uses 1-shot CLIP (ViT-B/16) over 300 epochs. Lines show means and shaded bands show one population standard deviation across seeds 1, 2, and 3.

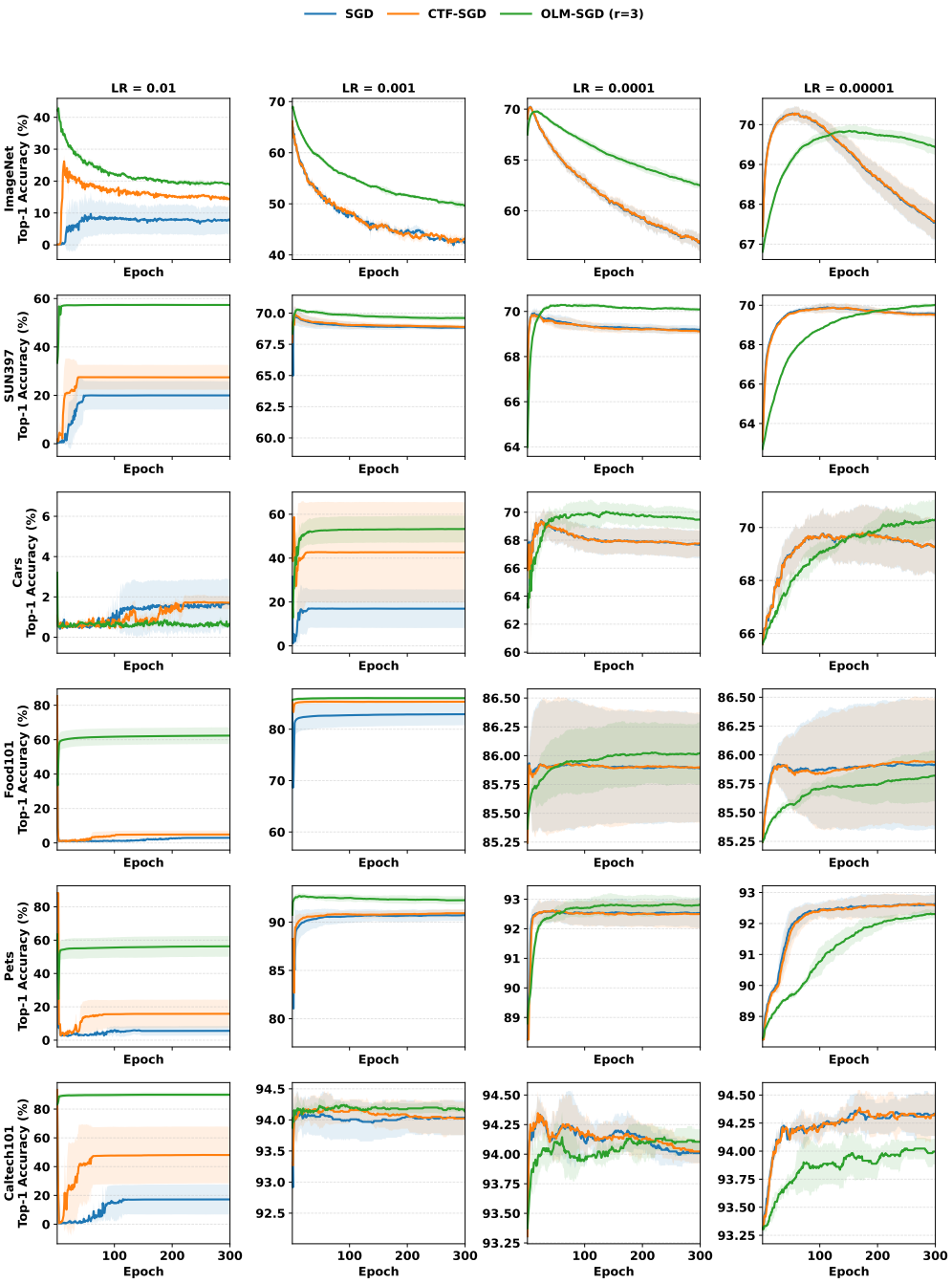

Figure 25: Test-accuracy trajectories for SGD, CTF-SGD, and OLM-SGD ($r = 3$), continuing Figure 24. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. ImageNet retains training augmentation under the evaluated protocol.

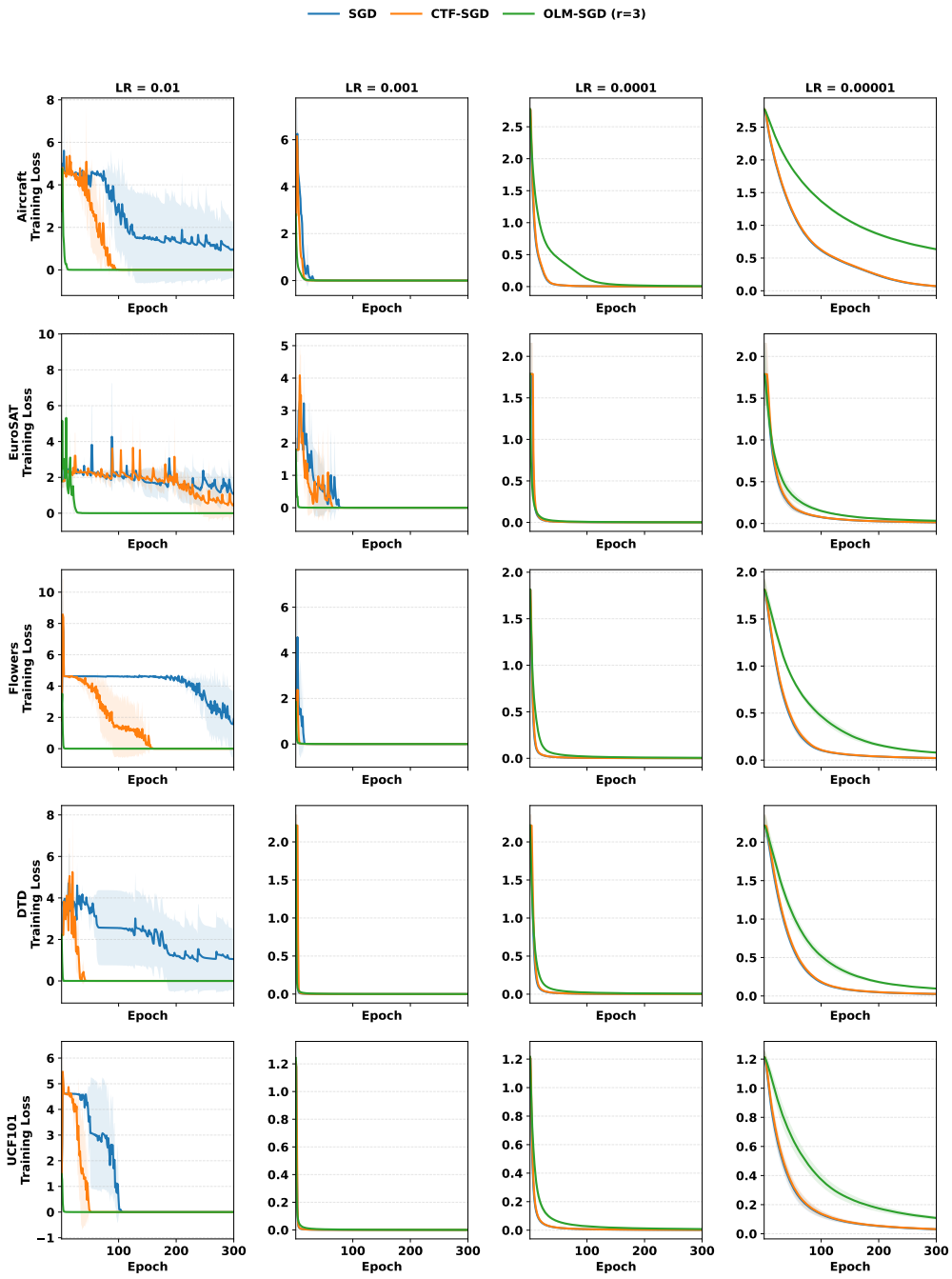

Figure 26: Training loss for the SGD, CTF-SGD, and OLM-SGD trajectories in Figure 24. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across the three seeds. These are support-set training losses, not test losses.

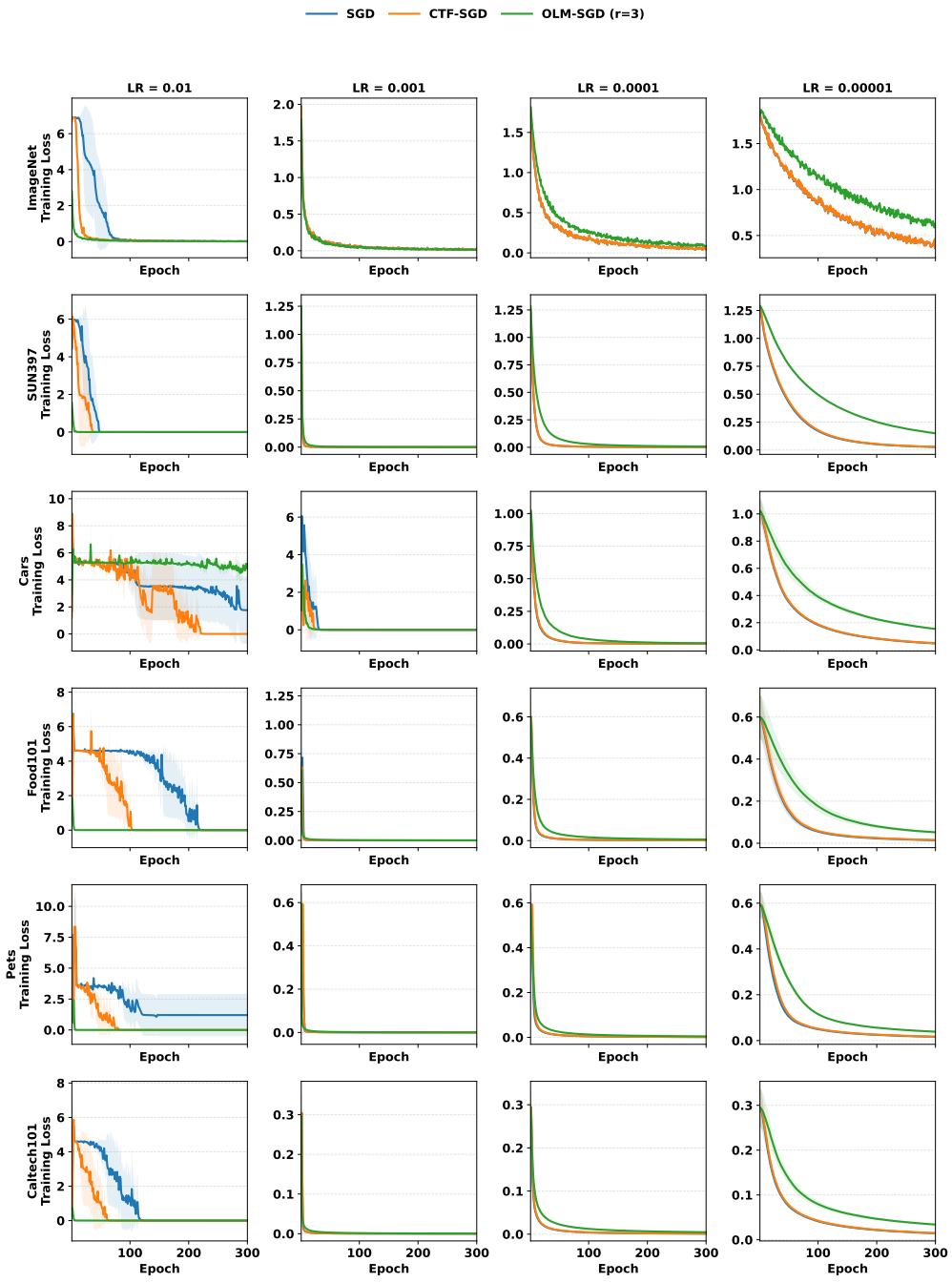

Figure 27: Training loss for the trajectories in Figure 25. Rows show ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines and bands use the same three-seed aggregation. All selected trajectories contain finite training losses through epoch 300. ImageNet retains training augmentation under the evaluated protocol.

H.3.4 BEST-TO-FINAL TEST-ACCURACY DROP

Smaller average drops at every learning rate. Tables 55–58 quantify accuracy retention. At 10^{-3}, OLM-SGD reduces the mean drop from CTF-SGD's 4.55 points to 2.22 points. At 10^{-4}, it reduces the drop from 1.80 to 0.95 points. These correspond to reductions of approximately 51% and 47%, respectively. At 10^{-5}, the drop decreases from 0.58 to 0.11 points. At 10^{-2}, it decreases from 30.45 to 3.45 points. In every case, the smaller mean drop accompanies higher mean final accuracy. The retention advantage is therefore not obtained merely by keeping the best and final checkpoints equally weak.

Consistent retention across datasets. At 10^{-4}, OLM-SGD has a smaller mean drop than CTF-SGD on all 11 datasets. ImageNet's drop decreases from 13.29 to 7.32 points. The reductions extend beyond ImageNet: Stanford Cars decreases from 1.83 to 0.72 points, and Aircraft decreases from 2.37 to 1.56 points. Excluding ImageNet, the ten-dataset average drops are 0.65 points for CTF-SGD and 0.31 for OLM-SGD. Hence, the average retention gain is not driven by a single target.

At 10^{-3}, Stanford Cars exposes pronounced seed-dependent degradation in CTF-SGD. Its drop is 19.03 ± 22.04 points, compared with 0.03 ± 0.02 for OLM-SGD. On Aircraft, the mean drop decreases from 5.09 to 2.20 points. At 10^{-5}, every OLM-SGD dataset-level mean drop is at most 0.43 points. This retention accompanies 71.75% mean final accuracy, rather than a failure to adapt.

Implication for the update design. CTF-SGD captures part of the improvement over SGD at larger learning rates. This supports the relevance of update-magnitude control. However, its global scalar shrinkage does not recover the full accuracy–retention trade-off of OLM-SGD. The matched trajectories show higher final accuracy, better transfer of the selected epoch, and smaller average degradation for the complete spectral merge. Together, these results support retaining low-rank filtering and polar reconstruction rather than replacing them with the tested direction-preserving scale control.

Table 55: Best-to-final test-accuracy drop (percentage points) at learning rate 0.01 (10^{-2}), for 1-shot CLIP(ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean $\pm$ population standard deviation over three seeds. Average summarizes the three per-seed unweighted averages over all 11 datasets.

Dataset	SGD	CTF-SGD	OLM-SGD ($r = 3$)
ImageNet	2.91 ± 2.42	12.82 ± 1.79	24.69 ± 0.49
Caltech101	76.09 ± 10.44	45.23 ± 19.73	0.07 ± 0.05
Oxford Pets	82.65 ± 2.92	72.44 ± 8.47	9.29 ± 6.93
Stanford Cars	0.29 ± 0.28	0.11 ± 0.02	2.83 ± 3.57
Flowers102	60.52 ± 7.07	43.48 ± 5.38	0.04 ± 0.06
Food101	82.30 ± 0.20	80.35 ± 2.26	0.00 ± 0.00
Aircraft	21.19 ± 1.04	20.65 ± 0.57	0.57 ± 0.39
SUN397	0.29 ± 0.34	0.17 ± 0.14	0.17 ± 0.17
DTD	34.75 ± 5.73	18.97 ± 5.22	0.12 ± 0.08
EuroSAT	16.09 ± 7.40	12.22 ± 9.96	0.17 ± 0.13
UCF101	39.20 ± 8.75	28.46 ± 4.71	0.03 ± 0.04
Average	37.84 ± 0.77	30.45 ± 0.33	3.45 ± 0.80

Table 56: Best-to-final test-accuracy drop (percentage points) at learning rate 0.001 (10^{-3}), for 1-shot CLIP(ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation over three seeds. Average summarizes the three per-seed unweighted averages over all 11 datasets.

Dataset	SGD	CTF-SGD	OLM-SGD ($r = 3$)
ImageNet	22.99 ± 0.31	22.91 ± 0.82	19.37 ± 0.63
Caltech101	0.18 ± 0.07	0.28 ± 0.18	0.20 ± 0.09
Oxford Pets	0.14 ± 0.19	0.11 ± 0.14	0.45 ± 0.15
Stanford Cars	14.59 ± 8.09	19.03 ± 22.04	0.03 ± 0.02
Flowers102	10.43 ± 14.53	0.04 ± 0.06	0.05 ± 0.02
Food101	2.48 ± 1.95	0.16 ± 0.11	0.05 ± 0.02
Aircraft	18.78 ± 1.53	5.09 ± 2.11	2.20 ± 0.39
SUN397	0.98 ± 0.55	1.02 ± 0.50	0.74 ± 0.18
DTD	0.28 ± 0.12	0.16 ± 0.12	0.14 ± 0.20
EuroSAT	0.04 ± 0.00	0.47 ± 0.41	0.84 ± 0.18
UCF101	0.48 ± 0.25	0.79 ± 0.39	0.29 ± 0.23
Average	6.49 ± 1.35	4.55 ± 2.23	2.22 ± 0.07

Table 57: Best-to-final test-accuracy drop (percentage points) at learning rate 0.0001 (10^{-4}), for 1-shot CLIP(ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation over three seeds. Average summarizes the three per-seed unweighted averages over all 11 datasets.

Dataset	SGD	CTF-SGD	OLM-SGD ($r = 3$)
ImageNet	13.36 ± 0.81	13.29 ± 0.85	7.32 ± 0.34
Caltech101	0.42 ± 0.18	0.38 ± 0.20	0.16 ± 0.10
Oxford Pets	0.23 ± 0.16	0.25 ± 0.18	0.07 ± 0.05
Stanford Cars	1.93 ± 0.63	1.83 ± 0.56	0.72 ± 0.13
Flowers102	0.08 ± 0.07	0.08 ± 0.09	0.04 ± 0.00
Food101	0.26 ± 0.21	0.27 ± 0.25	0.02 ± 0.02
Aircraft	2.24 ± 0.33	2.37 ± 0.26	1.56 ± 0.22
SUN397	0.80 ± 0.01	0.80 ± 0.05	0.28 ± 0.04
DTD	0.24 ± 0.17	0.26 ± 0.24	0.20 ± 0.17
EuroSAT	0.21 ± 0.23	0.12 ± 0.13	0.01 ± 0.02
UCF101	0.11 ± 0.08	0.19 ± 0.07	0.04 ± 0.04
Average	1.81 ± 0.12	1.80 ± 0.14	0.95 ± 0.06

Table 58: Best-to-final test-accuracy drop (percentage points) at learning rate 0.00001 (10^{-5}), for 1-shot CLIP(ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation over three seeds. Average summarizes the three per-seed unweighted averages over all 11 datasets.

Dataset	SGD	CTF-SGD	OLM-SGD ($r = 3$)
ImageNet	2.76 ± 0.27	2.73 ± 0.27	0.43 ± 0.06
Caltech101	0.07 ± 0.07	0.05 ± 0.05	0.07 ± 0.04
Oxford Pets	0.11 ± 0.14	0.16 ± 0.08	0.02 ± 0.03
Stanford Cars	0.69 ± 0.17	0.69 ± 0.15	0.19 ± 0.16
Flowers102	0.03 ± 0.04	0.01 ± 0.02	0.01 ± 0.02
Food101	0.22 ± 0.24	0.20 ± 0.24	0.06 ± 0.09
Aircraft	1.70 ± 0.04	1.85 ± 0.14	0.29 ± 0.21
SUN397	0.38 ± 0.10	0.40 ± 0.17	0.02 ± 0.00
DTD	0.12 ± 0.10	0.16 ± 0.15	0.08 ± 0.06
EuroSAT	0.06 ± 0.08	0.11 ± 0.12	0.00 ± 0.01
UCF101	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
Average	0.56 ± 0.05	0.58 ± 0.06	0.11 ± 0.00

H.4 Comparison against GaLore

GaLoRE tests whether low-rank processing of the minibatch mean gradient recovers the accuracy and stability of OLM-SGD. GaLoRE (Zhao et al., 2024) projects the aggregated gradient into a low-rank subspace. The evaluated implementation maintains Adam-style moments in that subspace and projects the update back to the original parameter space. It uses projection rank $r_{\mathrm{GaL}} = 2$ and refreshes the basis every ten optimizer steps.

H.4.1 Few-Shot Accuracy across Backbones and Shot Levels

Consistent gains in average accuracy. Tables 60–62 report the detailed results. Table 59 summarizes their Average columns. OLM-SGD exceeds GaLoRE in all 15 backbone–shot settings. On ViT-B/32, its gains at 1, 2, 4, 8, and 16 shots are 2.6, 2.3, 1.9, 2.3, and 2.0 points. On ViT-B/16, they are 2.6, 2.1, 2.2, 2.3, and 2.0 points. Averaging these displayed margins gives 2.22 and 2.24 points, respectively. The advantage therefore persists as the number of labeled examples increases.

On ViT-L/14, OLM-SGD also leads at every shot level. The margins range from 0.2 to 0.9 points. Averaging the five displayed margins gives 0.60 points. At 16 shots, it reaches 86.8%, compared with 85.9% for GaLoRE. The improvement is smaller on the largest backbone but remains positive throughout the evaluated shot range.

Table 59: Dataset-average few-shot accuracy (%) from Tables 60–62. We reproduce the displayed Average means. The final column subtracts GaLoRE from OLM-SGD, in percentage points. Bold marks the higher displayed average in each row.

Backbone	Shots	GaLoRE	OLM-SGD	OLM-SGD − GaLoRE
ViT-B/32	1	64.6	**67.2**	+2.6
	2	68.1	**70.4**	+2.3
	4	71.2	**73.1**	+1.9
	8	73.5	**75.8**	+2.3
	16	76.5	**78.5**	+2.0
ViT-B/16	1	69.4	**72.0**	+2.6
	2	73.0	**75.1**	+2.1
	4	75.6	**77.8**	+2.2
	8	78.0	**80.3**	+2.3
	16	80.6	**82.6**	+2.0
ViT-L/14	1	76.9	**77.4**	+0.5
	2	79.9	**80.1**	+0.2
	4	81.9	**82.5**	+0.6
	8	83.9	**84.7**	+0.8
	16	85.9	**86.8**	+0.9

The improvement extends across datasets. Using the displayed means, OLM-SGD strictly exceeds GaLoRE in $52/55$, $50/55$, and $36/55$ dataset–shot settings on ViT-B/32, ViT-B/16, and ViT-L/14. There are two additional ties on ViT-B/16 and three on ViT-L/14. Across the three backbones, OLM-SGD leads on ImageNet, Stanford Cars, Aircraft, and Food101 in all 15 settings per dataset. For 1-shot ViT-B/32, Stanford Cars improves from 56.0% to 62.6%, while ImageNet improves from 55.9% to 61.9%. For 16-shot ViT-B/16, Aircraft improves from 46.6% to 52.9%. These repeated gains show that the average improvement is not confined to one benchmark. GaLoRE remains competitive on EuroSAT, where it leads in $11/15$ settings, but this does not reverse the overall advantage on any backbone or shot level.

The fixed-schedule tables use 30 epochs for GaLoRE and 70 epochs for OLM-SGD, both at 10^{-4}. They measure adaptation under the reported schedules. The following 1-shot experiments match the learning rate and the 300-epoch horizon. At one shot, the randomized and class-unique samplers also have the same batching distribution, as discussed in the shared protocol.

Table 60: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. OLM-SGD is compared with GaLoRE. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. The Average column reports the mean with equal weight for the 11 datasets.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	GaLoRE	55.9 ± 0.6	64.6 ± 0.3	20.3 ± 1.4	67.1 ± 3.8	56.0 ± 1.1	77.3 ± 0.5	83.7 ± 2.4	75.3 ± 0.8	92.6 ± 0.7	48.1 ± 0.2	69.7 ± 0.2	64.6
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.7	50.7 ± 1.9	70.2 ± 0.4	67.2
2	GaLoRE	57.8 ± 0.2	65.9 ± 0.6	21.6 ± 0.3	78.2 ± 2.5	58.6 ± 0.5	78.3 ± 0.1	88.9 ± 0.5	80.8 ± 1.3	93.5 ± 0.3	53.7 ± 1.3	72.2 ± 0.4	68.1
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.8 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.3 ± 0.9	70.4
4	GaLoRE	58.7 ± 0.4	67.1 ± 0.3	26.1 ± 0.4	86.0 ± 1.6	62.8 ± 0.5	79.0 ± 0.3	89.9 ± 0.4	85.7 ± 0.8	94.4 ± 0.2	58.9 ± 0.3	74.9 ± 1.1	71.2
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.2 ± 0.6	81.3 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1
8	GaLoRE	60.3 ± 0.4	69.1 ± 0.2	30.8 ± 1.1	88.1 ± 2.0	67.4 ± 0.5	79.4 ± 0.2	90.5 ± 0.4	88.9 ± 0.2	94.8 ± 0.4	62.6 ± 0.2	77.0 ± 0.3	73.5
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.7 ± 0.1	91.2 ± 0.5	92.0 ± 0.1	95.2 ± 0.2	64.6 ± 0.7	78.5 ± 0.4	75.8
16	GaLoRE	63.0 ± 0.1	71.1 ± 0.1	37.9 ± 0.2	93.2 ± 0.7	72.0 ± 0.3	79.6 ± 0.1	90.6 ± 0.1	93.1 ± 0.5	95.2 ± 0.2	66.6 ± 1.0	79.5 ± 0.1	76.5
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5

Table 61: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. OLM-SGD is compared with GaLoRE. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. The Average column reports the mean with equal weight for the 11 datasets.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	GaLoRE	64.0 ± 0.2	67.8 ± 0.3	27.1 ± 0.8	68.6 ± 5.3	64.5 ± 1.0	83.5 ± 0.2	88.0 ± 1.5	81.6 ± 1.2	93.5 ± 0.2	50.8 ± 1.5	74.0 ± 1.1	69.4
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.4	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0
2	GaLoRE	64.3 ± 0.9	69.2 ± 0.5	31.0 ± 1.6	79.5 ± 2.2	68.2 ± 0.6	84.2 ± 0.3	92.0 ± 0.9	86.2 ± 0.4	94.5 ± 0.1	58.3 ± 1.4	76.0 ± 0.6	73.0
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.1 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1
4	GaLoRE	65.6 ± 0.4	70.1 ± 0.1	34.4 ± 0.8	86.1 ± 2.2	71.3 ± 0.5	85.3 ± 0.1	92.9 ± 0.2	90.2 ± 0.2	95.2 ± 0.2	62.0 ± 0.8	78.3 ± 1.1	75.6
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.1	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8
8	GaLoRE	66.5 ± 0.2	71.9 ± 0.4	39.9 ± 0.5	90.2 ± 0.4	75.1 ± 0.3	85.6 ± 0.3	93.1 ± 0.3	93.2 ± 0.1	95.6 ± 0.3	66.4 ± 0.5	80.7 ± 0.1	78.0
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.2	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3
16	GaLoRE	68.7 ± 0.1	73.7 ± 0.1	46.6 ± 0.4	93.5 ± 0.6	79.5 ± 0.6	85.8 ± 0.3	93.8 ± 0.2	96.0 ± 0.3	96.1 ± 0.4	70.7 ± 1.0	82.6 ± 0.2	80.6
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.1	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6

Table 62: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. OLM-SGD is compared with GaLoRE. Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. The Average column reports the mean with equal weight for the 11 datasets.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	GaLoRE	72.4 ± 0.5	73.7 ± 0.3	38.8 ± 1.7	75.8 ± 4.2	78.3 ± 0.2	89.1 ± 0.4	90.8 ± 1.8	89.3 ± 1.2	96.1 ± 0.6	60.2 ± 1.8	81.2 ± 0.8	76.9
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.5 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4
2	GaLoRE	73.1 ± 0.3	75.3 ± 0.3	41.5 ± 2.1	85.2 ± 1.1	80.5 ± 0.2	90.3 ± 0.1	94.0 ± 0.3	93.0 ± 0.9	96.8 ± 0.1	66.4 ± 0.6	83.0 ± 1.3	79.9
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1
4	GaLoRE	73.7 ± 0.2	75.9 ± 0.2	47.6 ± 0.9	87.1 ± 1.4	82.6 ± 0.5	90.9 ± 0.1	95.0 ± 0.2	96.0 ± 0.3	97.3 ± 0.2	68.8 ± 0.1	85.5 ± 0.6	81.9
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.2	91.3 ± 0.1	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.3	68.7 ± 0.9	84.8 ± 0.5	82.5
8	GaLoRE	74.5 ± 0.2	77.4 ± 0.6	54.1 ± 0.4	90.6 ± 1.3	85.2 ± 0.4	91.2 ± 0.2	95.5 ± 0.1	97.2 ± 0.2	97.6 ± 0.3	72.5 ± 0.7	86.7 ± 0.3	83.9
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7
16	GaLoRE	75.9 ± 0.1	78.7 ± 0.2	61.9 ± 1.0	94.2 ± 0.6	87.2 ± 0.2	91.4 ± 0.1	95.7 ± 0.2	98.3 ± 0.1	97.8 ± 0.1	76.0 ± 0.5	88.1 ± 0.3	85.9
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8

H.4.2 Cross-Dataset Transfer of the Selected Training Epoch

Higher transfer accuracy with a smaller stopping-time penalty. Figure 28 and Table 63 evaluate source-selected durations at 10^{-4}. OLM-SGD reaches 71.73% mean accuracy across the 110 off-diagonal transfers, compared with 66.72% for GaLoRE. Its mean gap to the target oracle is 0.91 points rather than 4.14. It remains within one point of the target oracle on $85/110$ pairs, compared with $37/110$ for GaLoRE.

The 5.01-point transfer advantage combines two effects. OLM-SGD's target-oracle average is 1.78 points higher. Its penalty for transferring the duration is another 3.23 points smaller. Thus, most of the transfer advantage comes from retaining more of the attainable accuracy under a source-selected stopping time. The benefit is not explained only by a better target-specific checkpoint.

Reducing both early- and late-epoch mismatch. GaLoRE selects epochs between 1 and 296, while OLM-SGD selects epochs between 18 and 292. Several GaLoRE sources select epochs 1–3. Transferring these durations to EuroSAT leaves accuracy at 42.21%, which is 26.56 points below its target oracle. Conversely, the late epoch selected by DTD reduces GaLoRE's ImageNet accuracy

Table 63: Epoch-transfer diagnostic for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. The target oracle averages each dataset's best seed-mean accuracy. Transfer accuracy and oracle gaps average the 110 off-diagonal source–target pairs. The final column counts pairs within one percentage point of the target oracle. These are test-guided diagnostics.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp
GaLoRE	70.86	66.72	4.14	37/110
OLM-SGD	**72.64**	**71.73**	**0.91**	**85/110**

to 46.97%. OLM-SGD's largest oracle gap across all transfers is 7.17 points. Its advantage reflects better accuracy retention across disparate stopping times.

Averaging the ten transferred durations for each target, OLM-SGD exceeds GaLoRE on all 11 datasets. The largest gains occur on EuroSAT (16.72 points), Flowers102 (8.73), and Stanford Cars (6.61). It improves over zero-shot CLIP on $103/110$ off-diagonal pairs, compared with $65/110$ for GaLoRE. All 100 transfers to targets other than ImageNet remain positive for OLM-SGD. On ImageNet, the largest decrease relative to zero-shot CLIP is reduced from 19.75 points to 4.11. These results use the test-guided diagnostic in the shared protocol. Only the duration is transferred, not weights or examples.

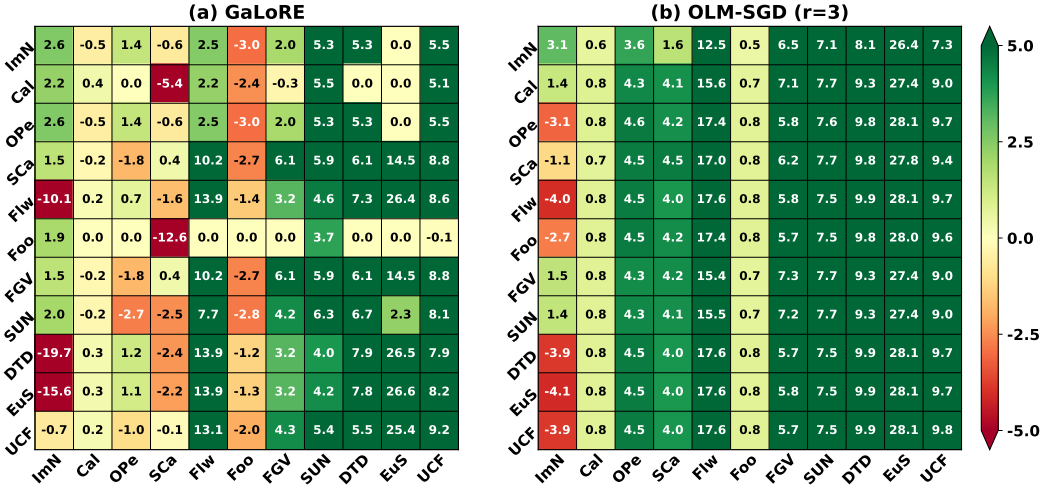

Figure 28: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Each source row selects the earliest epoch in 1–300 that maximizes mean *test* accuracy across seeds 1, 2, and 3. Each target column evaluates its independently trained trajectory at that epoch. Entries give the target accuracy change relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Colors saturate at ± 5 points, while annotations retain the rounded numerical values. OLM-SGD uses $r = 3$. Only the duration is transferred. Diagonal entries use target-specific oracle selection and are excluded from Table 63. This is a test-guided diagnostic, not validation-selected early stopping.

H.4.3 Long-Horizon Dynamics and Learning-Rate Sensitivity

A broader range of effective learning rates. Figures 29 and 30 show the test-accuracy trajectories. Table 64 summarizes their endpoints. At 10^{-3}, OLM-SGD reaches 69.23% mean final accuracy, compared with 36.82% for GaLoRE. It leads on all 11 datasets. Large differences occur on Flowers102 (86.26% versus 27.38%), Food101 (86.04% versus 31.02%), and Stanford Cars (53.31% versus 1.23%). At 10^{-4}, OLM-SGD again leads on every dataset. Its mean final accuracy is 71.75%, compared with 67.78% for GaLoRE.

Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of mean final accuracy is 2.52 points for OLM-SGD and 32.90 points for GaLoRE. Including 10^{-2} expands these ranges to 24.37 and 67.46 points. At this

most aggressive rate, OLM-SGD retains 47.38% mean final accuracy, whereas GaLoRE falls to 2.25%. The extreme rate still reduces OLM-SGD's accuracy substantially. Nevertheless, its relative advantage persists across the entire sweep.

Table 64: Long-horizon summary for 1-shot CLIP (ViT-B/16). Final accuracy is evaluated at epoch 300. Best-to-final drops are computed within each seed before aggregation. For both quantities, we first average the 11 datasets within each seed and then report mean $\pm$ population standard deviation across seeds. The last column gives the mean accuracy change from epoch 30 to epoch 300.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	GaLoRE	2.25 ± 0.32	44.24 ± 0.31	$+0.23$
	OLM-SGD	47.38 ± 0.43	3.45 ± 0.80	-0.18
10^{-3}	GaLoRE	36.82 ± 2.15	20.48 ± 1.71	$+1.98$
	OLM-SGD	69.23 ± 0.54	2.22 ± 0.07	-0.91
10^{-4}	GaLoRE	67.78 ± 0.55	3.26 ± 0.15	-1.57
	OLM-SGD	71.75 ± 0.26	0.95 ± 0.06	$+0.14$
10^{-5}	GaLoRE	69.72 ± 0.05	1.64 ± 0.05	-0.59
	OLM-SGD	71.75 ± 0.26	0.11 ± 0.00	$+3.85$

Preserving accuracy beyond the initial adaptation phase. At 10^{-4}, OLM-SGD's mean accuracy increases from 71.61% at epoch 30 to 71.75% at epoch 300. GaLoRE decreases from 69.35% to 67.78%. From epoch 70 onward, OLM-SGD loses only 0.27 points, compared with 1.17 for GaLoRE. Averaging the seed standard deviations across datasets and epochs also gives a smaller value: 0.57 points rather than 1.12. This combines better late-epoch accuracy retention with less seed variability at the evaluated rate.

At 10^{-5}, GaLoRE adapts faster initially. Its epoch-30 average is 70.31%, compared with 67.89% for OLM-SGD. Continuing to epoch 300 reverses the ranking. OLM-SGD gains 3.85 points and reaches 71.75%, whereas GaLoRE loses 0.59 points and finishes at 69.72%. OLM-SGD leads on eight datasets at the endpoint. The other three differences are each below 0.09 points. Thus, a slower initial improvement does not imply a weaker final model. OLM-SGD continues to benefit from training after GaLoRE's early advantage has diminished.

Low support loss does not ensure high test accuracy. Figures 31 and 32 separate the training-set fitting from generalization. At 10^{-4}, GaLoRE reaches a lower final training loss than OLM-SGD on all ten datasets except ImageNet. Yet, OLM-SGD has a higher final test accuracy on every dataset. On Stanford Cars, GaLoRE's final loss is approximately 9.41×10^{-5}, with 63.18% test accuracy. OLM-SGD retains 69.47% accuracy with a larger loss of approximately 5.54×10^{-3}.

The same distinction is more pronounced at 10^{-3}. On Flowers102, GaLoRE reaches a final loss of approximately 4.16×10^{-5} but only 27.38% test accuracy. OLM-SGD reaches 86.26% with a loss of approximately 4.91×10^{-4}. Thus, low-rank projection and rapid support fitting alone do not reproduce OLM-SGD's generalization behavior.

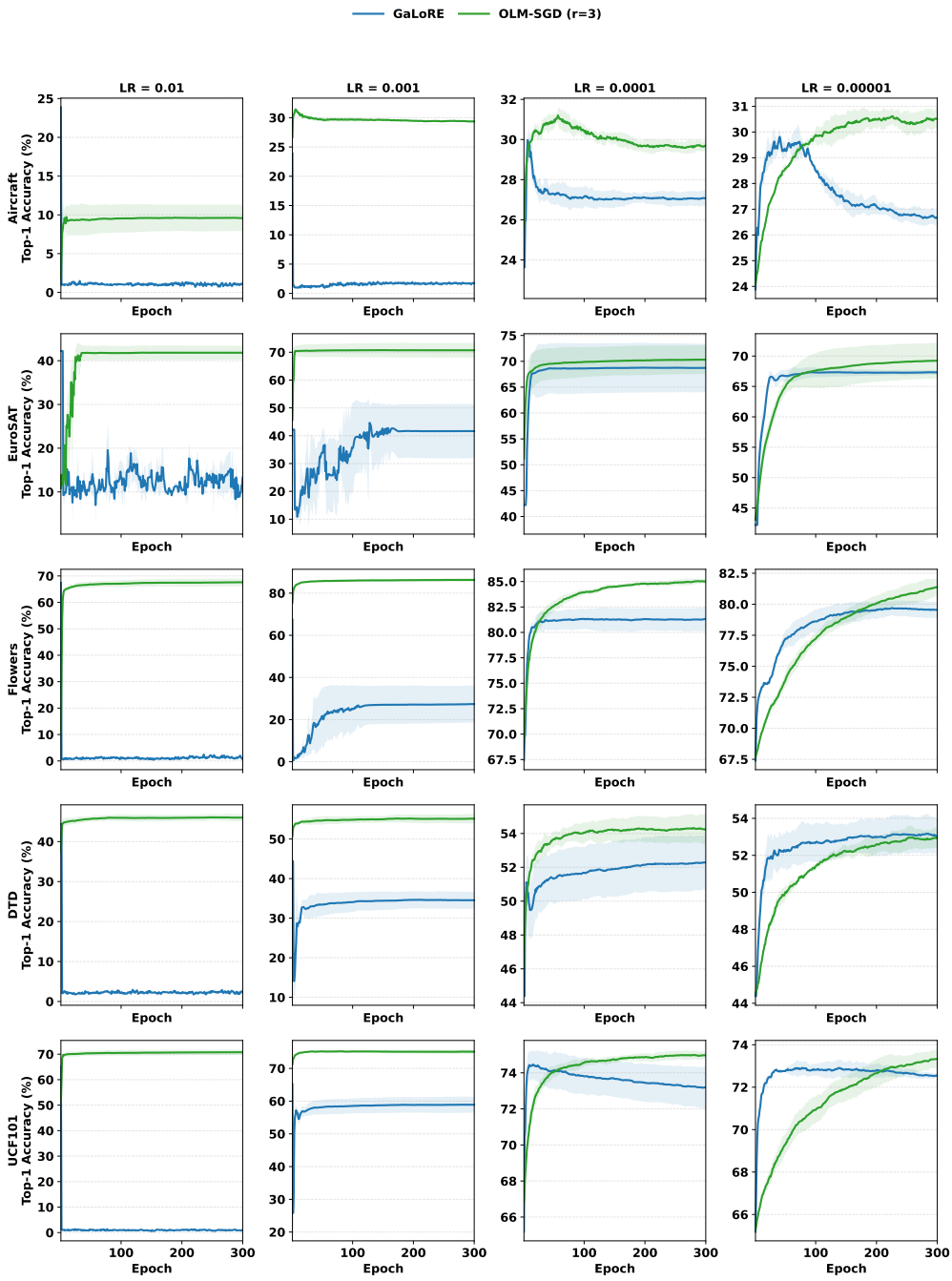

Figure 29: Test top-1 accuracy over 300 epochs of 1-shot CLIP (ViT-B/16) adaptation: GaLoRE and OLM-SGD ($r = 3$). Rows: Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns: learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show $\pm$ one population standard deviation across seeds 1, 2, and 3. All runs contain complete test-accuracy histories.

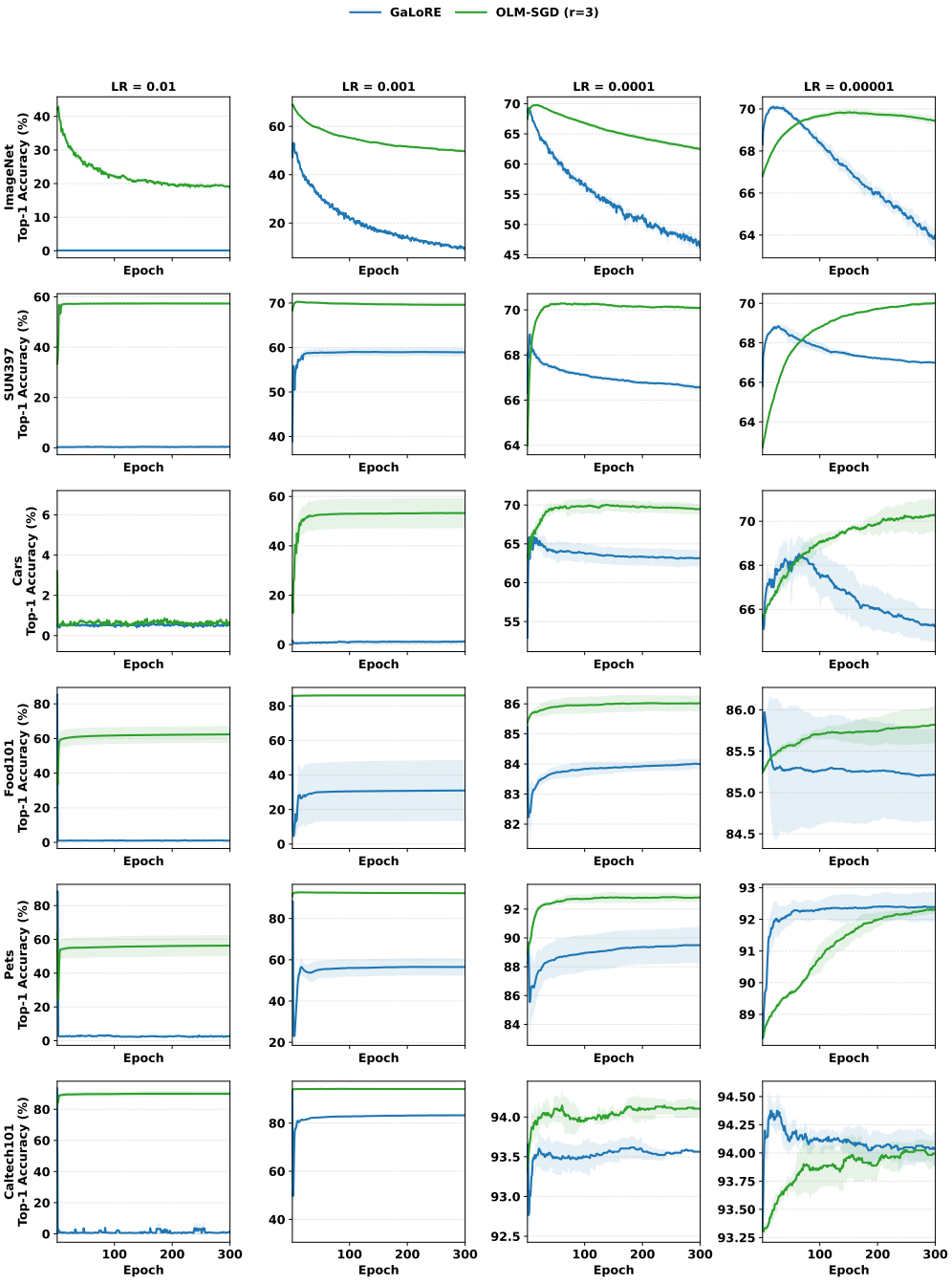

Figure 30: Test top-1 accuracy over 300 epochs of 1-shot CLIP (ViT-B/16) adaptation: GaLoRE and OLM-SGD ($r = 3$). Rows: ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns: learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show $\pm$ one population standard deviation across seeds 1, 2, and 3. All runs contain complete test-accuracy histories.

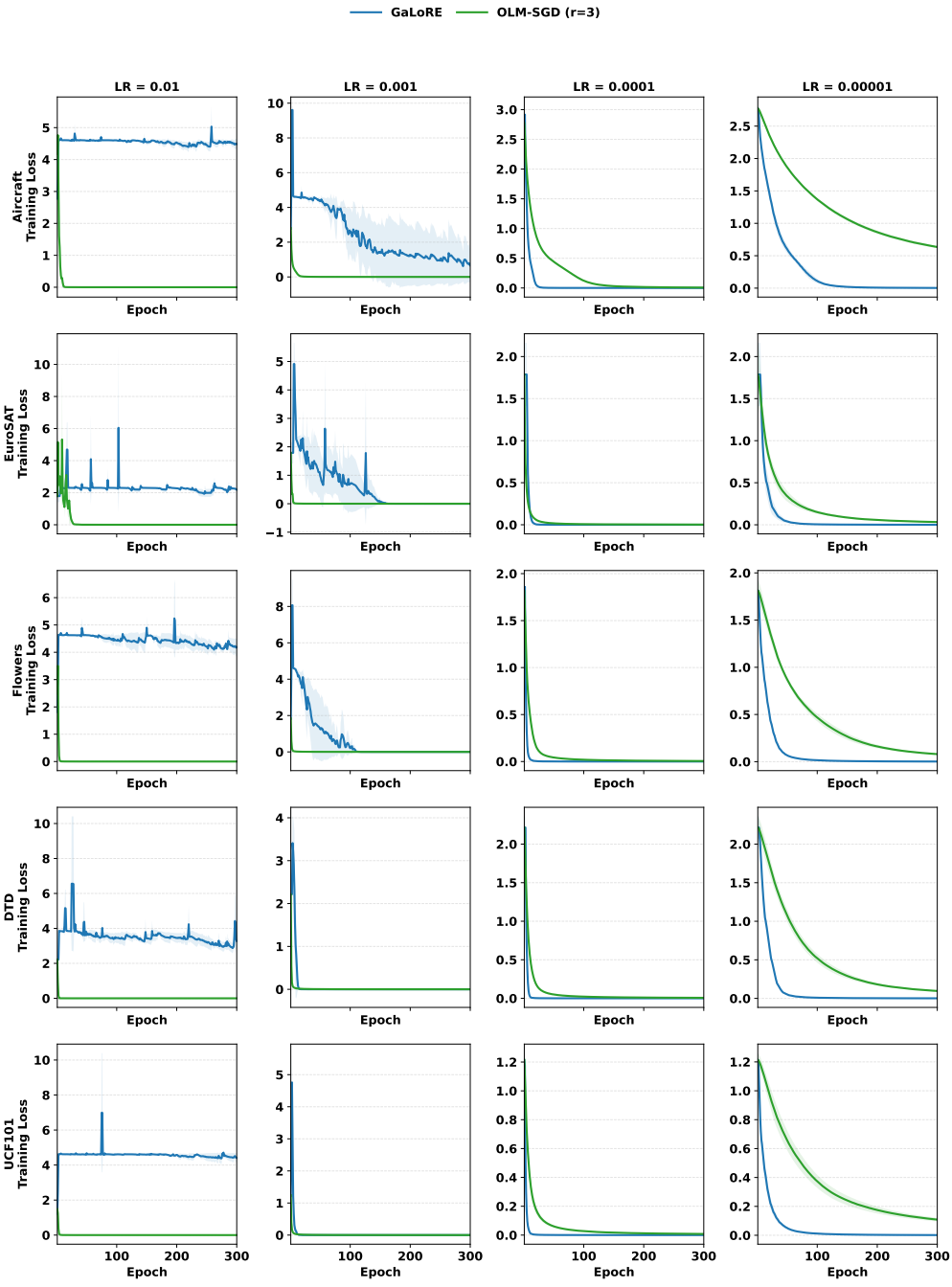

Figure 31: Training cross-entropy loss over 300 epochs of 1-shot CLIP (ViT-B/16) adaptation: GaLoRE and OLM-SGD ($r = 3$). Rows: Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns: learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show $\pm$ one population standard deviation across seeds 1, 2, and 3. These are training losses, not test losses. All three seeds have finite loss values at every epoch.

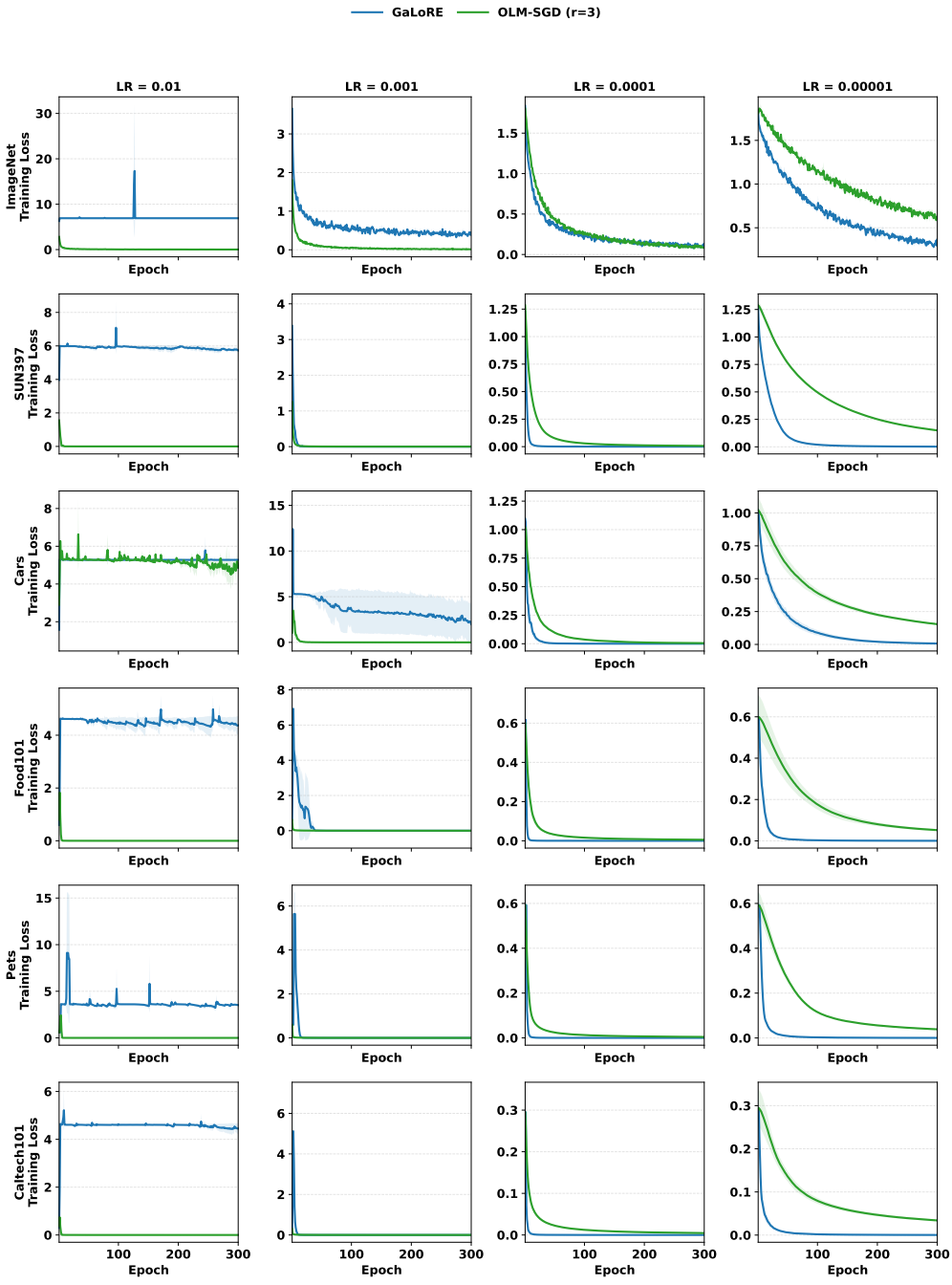

Figure 32: Training cross-entropy loss over 300 epochs of 1-shot CLIP (ViT-B/16) adaptation: GaLoRE and OLM-SGD ($r = 3$). Rows: ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns: learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show $\pm$ one population standard deviation across seeds 1, 2, and 3. These are training losses, not test losses. All three seeds have finite loss values at every epoch.

H.4.4 BEST-TO-FINAL TEST-ACCURACY DROP

Smaller degradation at every learning rate. Tables 65–68 quantify the distance from each seed's best checkpoint to epoch 300. OLM-SGD has the smallest dataset-average drop at all four learning rates. Its drops are 3.45, 2.22, 0.95, and 0.11 points at 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. GaLoRE's corresponding drops are 44.24, 20.48, 3.26, and 1.64 points. The advantage remains visible even at the smallest rate, where GaLoRE recovers strong absolute accuracy.

Retention gains are not confined to ImageNet. At 10^{-4}, OLM-SGD has a smaller mean drop on ten of the eleven datasets. The dataset-level mean drop is below one point on nine datasets, compared with five for GaLoRE. On ImageNet, the drop decreases from 23.14 to 7.32 points. Excluding ImageNet, the ten-dataset average still decreases from 1.27 to 0.31 points. Stanford Cars improves from 3.29 to 0.72, SUN397 from 2.45 to 0.28, and UCF101 from 1.44 to 0.04. DTD is the only exception in drop, at 0.20 points for OLM-SGD versus 0.04 for GaLoRE. Even there, OLM-SGD finishes at higher accuracy: 54.26% versus 52.29%.

At 10^{-3}, GaLoRE loses more than five points on nine datasets. OLM-SGD exceeds that drop only on ImageNet. Excluding ImageNet gives mean drops of 18.10 and 0.50 points, respectively. At 10^{-5}, OLM-SGD has a smaller drop on all 11 datasets, and no dataset-level mean exceeds 0.43 points. Its 0.11-point overall drop accompanies 71.75% final accuracy.Thus, the small drop reflects useful adaptation rather than uniformly weak checkpoints.

Accuracy and retention provide complementary evidence. At 10^{-2}, GaLoRE's ImageNet drop is only 0.07 points, but its final accuracy is 0.10%. A small drop can therefore occur when even the best recorded checkpoint has low accuracy. The final-accuracy and drop comparisons must be read together. Across the sweep, OLM-SGD combines higher mean final accuracy with smaller mean degradation. Together with the fixed-schedule and epoch-transfer results, this supports the benefit of structured individual-gradient merging over the evaluated low-rank projection baseline. The gain is not merely a better selected checkpoint. It also includes reduced sensitivity to the learning rate and training duration.

Table 65: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.01 (10^{-2}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	GaLoRE	OLM-SGD ($r = 3$)
ImageNet	0.07 ± 0.02	24.69 ± 0.49
Caltech101	92.25 ± 0.61	0.07 ± 0.05
Oxford Pets	85.41 ± 0.83	9.29 ± 6.93
Stanford Cars	0.15 ± 0.09	2.83 ± 3.57
Flowers102	66.31 ± 0.37	0.04 ± 0.06
Food101	84.26 ± 0.04	0.00 ± 0.00
Aircraft	22.76 ± 0.14	0.57 ± 0.39
SUN397	0.20 ± 0.03	0.17 ± 0.17
DTD	42.08 ± 0.38	0.12 ± 0.08
EuroSAT	28.91 ± 2.88	0.17 ± 0.13
UCF101	64.18 ± 0.08	0.03 ± 0.04
Average	44.24 ± 0.31	3.45 ± 0.80

Table 66: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.001 (10^{-3}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	GaLoRE	OLM-SGD ($r = 3$)
ImageNet	44.22 ± 0.75	19.37 ± 0.63
Caltech101	10.02 ± 0.06	0.20 ± 0.09
Oxford Pets	31.73 ± 4.14	0.45 ± 0.15
Stanford Cars	1.06 ± 1.09	0.03 ± 0.02
Flowers102	40.02 ± 8.80	0.05 ± 0.02
Food101	54.23 ± 17.66	0.05 ± 0.02
Aircraft	22.13 ± 0.22	2.20 ± 0.39
SUN397	0.15 ± 0.06	0.74 ± 0.18
DTD	9.85 ± 2.16	0.14 ± 0.20
EuroSAT	5.53 ± 5.25	0.84 ± 0.18
UCF101	6.29 ± 2.42	0.29 ± 0.23
Average	20.48 ± 1.71	2.22 ± 0.07

Table 67: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.0001 (10^{-4}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	GaLoRE	OLM-SGD ($r = 3$)
ImageNet	23.14 ± 0.48	7.32 ± 0.34
Caltech101	0.22 ± 0.10	0.16 ± 0.10
Oxford Pets	0.24 ± 0.17	0.07 ± 0.05
Stanford Cars	3.29 ± 0.30	0.72 ± 0.13
Flowers102	0.14 ± 0.08	0.04 ± 0.00
Food101	1.25 ± 0.18	0.02 ± 0.02
Aircraft	2.95 ± 0.21	1.56 ± 0.22
SUN397	2.45 ± 0.36	0.28 ± 0.04
DTD	0.04 ± 0.06	0.20 ± 0.17
EuroSAT	0.75 ± 0.38	0.01 ± 0.02
UCF101	1.44 ± 0.43	0.04 ± 0.04
Average	3.26 ± 0.15	0.95 ± 0.06

Table 68: Best-to-final test-accuracy drop (percentage points) for 1-shot CLIP (ViT-B/16) at learning rate 0.00001 (10^{-5}). Lower is better. For each seed, the drop is its highest test accuracy over epochs 1–300 minus its accuracy at epoch 300. All runs have complete test-accuracy histories through epoch 300. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. The Average row first averages the 11 dataset drops within each seed, then reports the mean and standard deviation across seeds.

Dataset	GaLoRE	OLM-SGD ($r = 3$)
ImageNet	6.20 ± 0.30	0.43 ± 0.06
Caltech101	0.46 ± 0.32	0.07 ± 0.04
Oxford Pets	0.08 ± 0.08	0.02 ± 0.03
Stanford Cars	3.41 ± 0.37	0.19 ± 0.16
Flowers102	0.20 ± 0.06	0.01 ± 0.02
Food101	0.85 ± 0.48	0.06 ± 0.09
Aircraft	3.57 ± 0.63	0.29 ± 0.21
SUN397	1.90 ± 0.13	0.02 ± 0.00
DTD	0.33 ± 0.07	0.08 ± 0.06
EuroSAT	0.47 ± 0.31	0.00 ± 0.01
UCF101	0.51 ± 0.25	0.01 ± 0.01
Average	1.64 ± 0.05	0.11 ± 0.00

H.5 Comparison against OLM-SGD with Class-Mean Gradients

This comparison tests whether averaging the support gradients within each class improves the contribution supplied to the spectral merge. OLM-SGD uses one example per selected class and processes different shots in successive updates. The class-mean variant instead computes the mean gradient of all K support examples from each selected class. It then applies the same tensor-wise truncation and polar reconstruction, as specified in Algorithm 9. Both methods use 70 epochs at learning rate 10^{-4} with gradient rank $r = 3$. They adapt the same parameters in both CLIP encoders. The class capacities are matched at $B_{\text{cls}} = B = \min\{64, C\}$.

H.5.1 Few-Shot Accuracy across Backbones and Shot Levels

One-shot equivalence. Tables 70–72 report the results across 11 datasets and seeds 1, 2, and 3. At $K = 1$, the class mean contains a single gradient. The two methods reduce to the same update under matched class groups and decomposition choices. Accordingly, all 33 dataset–backbone entries have identical displayed means and standard deviations. We focus the comparison on $K > 1$, where averaging within a class changes the input to the merge.

The average advantage grows with the number of shots. Table 69 summarizes the displayed Average columns. OLM-SGD leads in 10 of the 12 backbone–shot settings with $K > 1$. The two remaining settings are ties at the displayed precision. These occur at 2 shots on ViT-B/32 and ViT-B/16. At 4, 8, and 16 shots, the gains on ViT-B/32 are 0.2, 0.8, and 1.9 points. On ViT-B/16, they are 0.6, 1.2, and 2.2 points. Thus, the advantage increases as more labeled examples become available.

The separation is largest on ViT-L/14. Its margins increase from 0.5 points at 2 shots to 1.2, 2.1, and 3.4 points at 4, 8, and 16 shots. At 16 shots, OLM-SGD reaches 86.8%, compared with 83.4% for the class-mean variant. Thus, a larger backbone does not eliminate the benefit of processing the shots separately.

Table 69: Dataset-average few-shot accuracy (%) from Tables 70–72. Class-Mean denotes OLM-SGD (Class-Mean Gradients). We reproduce the displayed Average means. Gain subtracts Class-Mean from OLM-SGD, in percentage points. Bold marks a strictly higher displayed mean. The 1-shot rows coincide by construction.

Backbone	Shots	Class-Mean	OLM-SGD	Gain (pp)
ViT-B/32	1	67.2	67.2	0.0
	2	70.4	70.4	0.0
	4	72.9	**73.1**	+0.2
	8	75.0	**75.8**	+0.8
	16	76.6	**78.5**	+1.9
ViT-B/16	1	72.0	72.0	0.0
	2	75.1	75.1	0.0
	4	77.2	**77.8**	+0.6
	8	79.1	**80.3**	+1.2
	16	80.4	**82.6**	+2.2
ViT-L/14	1	77.4	77.4	0.0
	2	79.6	**80.1**	+0.5
	4	81.3	**82.5**	+1.2
	8	82.6	**84.7**	+2.1
	16	83.4	**86.8**	+3.4

Consistent gains across datasets. Using the displayed means and excluding the identical 1-shot entries, OLM-SGD strictly exceeds the class-mean variant in $28/44$, $35/44$, and $41/44$ dataset–shot settings on ViT-B/32, ViT-B/16, and ViT-L/14. There are two additional ties on ViT-B/32 and one on ViT-B/16. In total, OLM-SGD leads in $104/132$ settings. It improves EuroSAT, Flowers102, Caltech101, DTD, and UCF101 in all 12 settings per dataset. The average advantage is supported by repeated gains across multiple tasks.

ImageNet favors the class-mean variant in 11 of the 12 settings with $K > 1$. At 16 shots, its margins on ViT-B/32 and ViT-B/16 are 4.0 and 2.4 points. These results are retained in the 11-dataset averages, which still favor OLM-SGD by 1.9 and 2.2 points. On ViT-L/14 at 16 shots, OLM-SGD also improves ImageNet by 0.9 points and leads on all 11 datasets.

Larger improvements on fine-grained recognition and texture classification. Aircraft shows the largest separation. At 16 shots, accuracy increases from 33.5% to 41.4% on ViT-B/32 and from 43.0% to 52.9% on ViT-B/16. On ViT-L/14, it increases from $49.7 \pm 0.4\%$ to $63.1 \pm 0.1\%$. The corresponding gains are 7.9, 9.9, and 13.4 points. Stanford Cars also improves by 5.0, 4.9, and 4.7 points at 16 shots. Flowers102 gains 3.8, 2.8, and 2.8 points. The trend extends beyond these fine-grained tasks. DTD gains 2.7, 3.9, and 5.0 points, while UCF101 gains 3.4, 3.1, and 4.0 points. These improvements persist across all three backbone scales.

Additional support examples yield larger accuracy gains. Both methods improve when the support set grows. However, OLM-SGD gains more from increasing the shot count under the stated schedules. From 1 to 16 shots, its average accuracy increases by 11.3, 10.6, and 9.4 points on the three backbones. The class-mean variant gains 9.4, 8.4, and 6.0 points. The difference remains visible from 8 to 16 shots alone. OLM-SGD gains 2.7, 2.3, and 2.1 points, compared with 1.6, 1.3, and 0.8 points. Thus, the widening gap is not caused by a decline in class-mean accuracy. It reflects larger improvements from additional labeled examples with OLM-SGD.

Interpretation. Class-mean merging retains the spectral reconstruction but changes the gradient information entering it. Averaging all shots first changes the singular subspaces and retained energies before orthogonalization. Since truncation and polar reconstruction are nonlinear, this operation is not equivalent to processing the individual gradients across updates. OLM-SGD allows each shot to contribute its own retained components rather than compressing the class into one mean gradient.

The comparison also preserves the update schedules defined in Appendix C. Both methods visit all CK support examples per epoch. OLM-SGD makes $K\lceil C/B \rceil$ updates, whereas the class-mean variant makes $\lceil C/B \rceil$. Equal epochs therefore match support-set exposure rather than optimization steps. Under this protocol, the results favor processing individual shots through successive spectral updates over merging complete class means. This supports the proposed combination of class-unique minibatching and individual-gradient merging, especially when several labeled examples are available per class.

Table 70: Few-shot classification on 11 datasets with the CLIP(ViT-B/32) backbone. OLM-SGD is compared with OLM-SGD (Class-Mean Gradients). Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average reports the mean across seeds after averaging the 11 datasets within each seed with equal weight. Its standard deviation is not displayed.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.0	62.1	18.1	35.8	60.1	78.9	85.1	64.0	91.4	43.0	61.6	60.2
1	OLM-SGD (Class-Mean Gradients)	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.7	50.7 ± 1.9	70.2 ± 0.4	67.2
	OLM-SGD	61.9 ± 0.3	68.0 ± 0.1	22.1 ± 0.4	62.5 ± 4.5	62.6 ± 0.3	80.3 ± 0.1	88.8 ± 0.3	79.0 ± 0.6	93.4 ± 0.7	50.7 ± 1.9	70.2 ± 0.4	67.2
2	OLM-SGD (Class-Mean Gradients)	63.9 ± 0.2	69.4 ± 0.2	24.8 ± 0.4	76.0 ± 1.5	64.9 ± 0.4	81.0 ± 0.1	89.6 ± 0.5	82.6 ± 1.1	93.9 ± 0.1	55.5 ± 1.2	72.8 ± 1.0	70.4
	OLM-SGD	61.8 ± 0.2	69.3 ± 0.2	24.9 ± 0.2	76.6 ± 1.5	64.2 ± 0.1	81.0 ± 0.1	89.4 ± 0.7	83.8 ± 0.7	94.2 ± 0.1	56.1 ± 1.0	73.3 ± 0.9	70.4
4	OLM-SGD (Class-Mean Gradients)	66.2 ± 0.1	71.2 ± 0.1	27.7 ± 0.4	80.9 ± 1.8	67.5 ± 0.8	81.3 ± 0.1	90.5 ± 0.2	87.1 ± 0.4	94.6 ± 0.2	59.7 ± 0.4	75.1 ± 0.2	72.9
	OLM-SGD	62.0 ± 0.3	70.8 ± 0.2	28.5 ± 0.6	81.9 ± 2.0	67.2 ± 0.6	81.3 ± 0.1	90.7 ± 0.3	89.6 ± 0.3	94.8 ± 0.2	61.1 ± 0.1	76.4 ± 0.5	73.1
8	OLM-SGD (Class-Mean Gradients)	67.4 ± 0.1	72.3 ± 0.1	30.8 ± 0.3	88.5 ± 0.8	70.0 ± 0.4	81.8 ± 0.1	91.4 ± 0.3	89.1 ± 0.5	94.8 ± 0.2	62.7 ± 0.7	76.8 ± 0.1	75.0
	OLM-SGD	62.5 ± 0.3	72.2 ± 0.1	34.3 ± 0.7	90.0 ± 0.1	71.8 ± 0.1	81.7 ± 0.1	91.2 ± 0.5	92.0 ± 0.1	95.2 ± 0.2	64.6 ± 0.7	78.5 ± 0.4	75.8
16	OLM-SGD (Class-Mean Gradients)	68.4 ± 0.1	73.3 ± 0.1	33.5 ± 0.3	91.5 ± 0.5	71.5 ± 0.3	82.3 ± 0.1	91.5 ± 0.2	91.3 ± 0.2	95.2 ± 0.3	66.1 ± 0.2	77.7 ± 0.1	76.6
	OLM-SGD	64.4 ± 0.1	73.5 ± 0.1	41.4 ± 0.1	93.3 ± 0.3	76.5 ± 0.4	81.9 ± 0.1	91.4 ± 0.4	95.1 ± 0.4	95.7 ± 0.2	68.8 ± 0.4	81.1 ± 0.4	78.5

Table 71: Few-shot classification on 11 datasets with the CLIP(ViT-B/16) backbone. OLM-SGD is compared with OLM-SGD (Class-Mean Gradients). Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average reports the mean across seeds after averaging the 11 datasets within each seed with equal weight. Its standard deviation is not displayed.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	66.7	62.6	23.9	42.2	65.5	85.2	88.2	67.4	93.3	44.4	65.2	64.1
1	OLM-SGD (Class-Mean Gradients)	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.4	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0
	OLM-SGD	67.8 ± 0.3	70.3 ± 0.0	30.7 ± 0.5	69.7 ± 3.0	69.7 ± 1.4	86.0 ± 0.2	92.7 ± 0.3	83.3 ± 0.3	94.0 ± 0.1	53.9 ± 0.5	74.4 ± 0.3	72.0
2	OLM-SGD (Class-Mean Gradients)	69.5 ± 0.2	72.1 ± 0.1	33.9 ± 1.7	79.0 ± 1.6	72.7 ± 0.2	86.2 ± 0.2	92.5 ± 0.2	88.7 ± 0.7	94.4 ± 0.2	59.9 ± 0.9	76.8 ± 0.3	75.1
	OLM-SGD	67.8 ± 0.3	72.0 ± 0.1	33.1 ± 1.1	79.1 ± 1.7	72.6 ± 0.4	86.3 ± 0.2	92.6 ± 0.2	90.0 ± 0.4	94.5 ± 0.3	60.6 ± 0.6	77.1 ± 0.3	75.1
4	OLM-SGD (Class-Mean Gradients)	71.2 ± 0.0	73.7 ± 0.1	36.9 ± 0.8	84.0 ± 1.3	75.6 ± 0.1	86.7 ± 0.0	93.5 ± 0.2	91.7 ± 0.3	94.9 ± 0.1	62.9 ± 0.3	78.3 ± 0.7	77.2
	OLM-SGD	68.6 ± 0.1	73.4 ± 0.0	38.1 ± 0.6	84.8 ± 1.2	76.5 ± 0.2	86.8 ± 0.1	93.7 ± 0.1	93.6 ± 0.4	95.2 ± 0.2	65.0 ± 0.8	79.8 ± 0.5	77.8
8	OLM-SGD (Class-Mean Gradients)	72.4 ± 0.1	74.8 ± 0.2	40.7 ± 0.3	90.0 ± 0.2	77.0 ± 0.5	87.0 ± 0.1	93.3 ± 0.3	93.3 ± 0.2	95.3 ± 0.1	66.0 ± 0.8	79.8 ± 0.3	79.1
	OLM-SGD	69.4 ± 0.3	75.0 ± 0.2	44.9 ± 0.8	90.8 ± 0.4	79.3 ± 0.4	87.1 ± 0.1	93.8 ± 0.4	95.9 ± 0.2	95.7 ± 0.1	68.8 ± 0.6	82.3 ± 0.1	80.3
16	OLM-SGD (Class-Mean Gradients)	73.2 ± 0.1	75.7 ± 0.1	43.0 ± 0.4	92.2 ± 0.6	78.5 ± 0.3	87.5 ± 0.1	93.9 ± 0.1	94.8 ± 0.1	95.6 ± 0.0	68.8 ± 0.6	81.2 ± 0.1	80.4
	OLM-SGD	70.8 ± 0.1	76.0 ± 0.1	52.9 ± 0.6	93.2 ± 0.4	83.4 ± 0.2	87.5 ± 0.1	94.1 ± 0.1	97.6 ± 0.1	96.3 ± 0.2	72.7 ± 0.3	84.3 ± 0.1	82.6

Table 72: Few-shot classification on 11 datasets with the CLIP(ViT-L/14) backbone. OLM-SGD is compared with OLM-SGD (Class-Mean Gradients). Entries are top-1 accuracy (%) as mean ± population standard deviation over seeds 1, 2, and 3. Average reports the mean across seeds after averaging the 11 datasets within each seed with equal weight. Its standard deviation is not displayed.

Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	73.5	67.7	30.0	55.1	76.8	90.3	93.1	76.2	95.2	52.4	73.7	71.3
1	OLM-SGD (Class-Mean Gradients)	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.5 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4
	OLM-SGD	75.9 ± 0.1	73.6 ± 0.1	39.6 ± 1.4	71.5 ± 1.3	80.6 ± 0.1	90.6 ± 0.1	94.4 ± 0.2	88.5 ± 0.4	95.7 ± 0.6	60.0 ± 0.2	80.8 ± 0.7	77.4
2	OLM-SGD (Class-Mean Gradients)	76.9 ± 0.0	75.1 ± 0.0	43.0 ± 1.1	78.9 ± 1.8	81.8 ± 0.1	90.8 ± 0.2	94.5 ± 0.1	92.2 ± 0.6	96.2 ± 0.3	64.6 ± 0.5	81.8 ± 0.9	79.6
	OLM-SGD	76.6 ± 0.2	75.6 ± 0.1	43.6 ± 1.1	79.6 ± 1.5	82.6 ± 0.4	90.9 ± 0.2	94.7 ± 0.1	93.3 ± 0.4	96.5 ± 0.3	65.2 ± 0.8	82.4 ± 0.8	80.1
4	OLM-SGD (Class-Mean Gradients)	77.7 ± 0.1	76.1 ± 0.1	46.5 ± 0.3	84.3 ± 2.9	83.0 ± 0.3	91.0 ± 0.1	94.8 ± 0.2	94.1 ± 0.3	96.7 ± 0.5	66.8 ± 0.7	83.4 ± 0.4	81.3
	OLM-SGD	77.3 ± 0.0	77.2 ± 0.1	49.2 ± 0.6	85.4 ± 2.7	84.4 ± 0.2	91.3 ± 0.1	95.1 ± 0.2	96.6 ± 0.1	97.4 ± 0.3	68.7 ± 0.9	84.8 ± 0.5	82.5
8	OLM-SGD (Class-Mean Gradients)	78.0 ± 0.1	76.9 ± 0.2	48.6 ± 0.2	89.5 ± 1.0	83.9 ± 0.1	91.4 ± 0.0	94.9 ± 0.4	95.1 ± 0.3	96.9 ± 0.2	69.1 ± 0.7	84.5 ± 0.4	82.6
	OLM-SGD	77.8 ± 0.1	78.8 ± 0.4	55.8 ± 0.7	90.1 ± 0.9	87.2 ± 0.4	91.7 ± 0.1	95.5 ± 0.2	97.8 ± 0.2	97.5 ± 0.4	72.8 ± 0.8	86.7 ± 0.3	84.7
16	OLM-SGD (Class-Mean Gradients)	78.3 ± 0.1	77.4 ± 0.1	49.7 ± 0.4	92.4 ± 0.8	84.4 ± 0.1	91.4 ± 0.1	94.9 ± 0.0	95.9 ± 0.1	97.1 ± 0.1	71.3 ± 0.6	84.7 ± 0.4	83.4
	OLM-SGD	79.2 ± 0.1	80.0 ± 0.1	63.1 ± 0.1	93.7 ± 0.6	89.1 ± 0.2	91.9 ± 0.1	95.9 ± 0.1	98.7 ± 0.1	97.8 ± 0.1	76.3 ± 0.3	88.7 ± 0.4	86.8

H.6 Comparison against OLM-SGD without Polar Reconstruction

This ablation tests whether low-rank filtering alone explains the benefits of OLM-SGD. It retains class-unique minibatching and the same rank-r_W approximation of each individual gradient. It then averages the truncated matrices directly. It omits the two polar transformations while preserving the retained singular-value coefficients and the single division by the actual batch size B_t. Algorithm 8 gives the complete procedure. This appendix reports the ten non-ImageNet datasets for which the paired comparison is available. The fixed-schedule tables cover ViT-B/32 and ViT-B/16 with $K \in \{1, 2, 4, 8, 16\}$. Both methods use 70 epochs at 10^{-4} with $r = 3$. The long-horizon experiments use 1-shot ViT-B/16 adaptation with the same rank.

H.6.1 Few-Shot Accuracy across Backbones and Shot Levels

Higher average accuracy under matched schedules. Tables 74 and 75 report the detailed results. Table 73 summarizes their Average columns. OLM-SGD leads in all ten backbone–shot settings. On ViT-B/32, its gains at 1, 2, 4, 8, and 16 shots are 0.25, 0.51, 0.49, 0.72, and 0.83 points. On ViT-B/16, they are 0.51, 0.46, 0.46, 0.73, and 0.80 points. Averaging the five displayed margins gives 0.56 and 0.59 points, respectively. These gains are obtained with the same epoch budget and class-unique batching rule. Thus, they reflect the reconstruction stage rather than additional support-set exposure.

Table 73: Dataset-average few-shot accuracy (%) from Tables 74 and 75. Both methods use the same 70-epoch schedule at 10^{-4} with $r = 3$. The averages give equal weight to the ten non-ImageNet datasets. We reproduce the displayed Average means. Gain subtracts the result without polar reconstruction from OLM-SGD, in percentage points. Bold marks the higher displayed mean.

Backbone	Shots	Without polar reconstruction	OLM-SGD	Gain (pp)
ViT-B/32	1	67.49	**67.74**	+0.25
	2	70.78	**71.29**	+0.51
	4	73.72	**74.21**	+0.49
	8	76.42	**77.14**	+0.72
	16	79.04	**79.87**	+0.83
ViT-B/16	1	71.94	**72.45**	+0.51
	2	75.33	**75.79**	+0.46
	4	78.24	**78.70**	+0.46
	8	80.62	**81.35**	+0.73
	16	83.01	**83.81**	+0.80

Consistent gains across datasets. Using the displayed means, OLM-SGD strictly improves $39/50$ dataset–shot settings on ViT-B/32 and $41/50$ on ViT-B/16. SUN397, Aircraft, Stanford Cars, Food101, and DTD improve in all ten settings per dataset across the two backbones. For example, 16-shot ViT-B/16 accuracy on Aircraft increases from 50.33% to 52.86%. On Stanford Cars with 1-shot ViT-B/32, accuracy increases from 60.08% to 62.56%. These gains show that the common low-rank filtering step is not sufficient to reproduce the accuracy of the polar merge.

The advantage becomes more consistent at larger shot counts. At 16 shots, OLM-SGD leads on all ten datasets with both backbones. It also leads on every dataset at 8 shots with ViT-B/16. Individual low-shot settings can favor direct reconstruction. For example, 1-shot ViT-B/32 accuracy on EuroSAT is 65.05% without polar reconstruction and 62.53% with it. Nevertheless, the dataset averages favor OLM-SGD at every shot level, and its largest average gains occur at 8 and 16 shots.

Table 74: Few-shot classification on the 10 non-ImageNet datasets with the CLIP(ViT-B/32) backbone. OLM-SGD is compared with OLM-SGD (without polar reconstruction). Both methods use 70 epochs at 10^{-4} with $r = 3$. Entries are top-1 accuracy (%) as mean $\pm$ population standard deviation over seeds 1, 2, and 3. Average reports the mean across seeds after averaging the 10 datasets within each seed with equal weight. Its standard deviation is not displayed.

Shots	Method	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.06	18.06	35.81	60.14	78.86	85.06	64.03	91.40	42.97	61.64	60.00
1	OLM-SGD (without polar reconstruction)	66.61 ± 0.15	21.54 ± 0.68	65.05 ± 2.02	60.08 ± 0.52	79.99 ± 0.30	89.08 ± 0.59	78.59 ± 1.08	93.67 ± 0.66	49.29 ± 0.84	71.01 ± 0.32	67.49
	OLM-SGD	67.97 ± 0.09	22.08 ± 0.44	62.53 ± 4.48	62.56 ± 0.33	80.26 ± 0.12	88.76 ± 0.32	79.04 ± 0.63	93.36 ± 0.65	50.69 ± 1.92	70.15 ± 0.37	67.74
2	OLM-SGD (without polar reconstruction)	67.75 ± 0.35	23.85 ± 0.31	78.37 ± 2.64	61.75 ± 0.21	80.26 ± 0.21	89.49 ± 0.66	83.12 ± 0.61	94.52 ± 0.20	55.04 ± 1.68	73.62 ± 0.41	70.78
	OLM-SGD	69.26 ± 0.17	24.94 ± 0.20	76.60 ± 1.53	64.15 ± 0.07	80.97 ± 0.10	89.43 ± 0.67	83.85 ± 0.73	94.20 ± 0.14	56.15 ± 0.99	73.35 ± 0.89	71.29
4	OLM-SGD (without polar reconstruction)	69.12 ± 0.26	28.18 ± 0.58	84.12 ± 1.17	65.20 ± 0.58	80.48 ± 0.18	90.60 ± 0.41	88.25 ± 0.71	94.92 ± 0.08	60.15 ± 0.12	76.17 ± 0.91	73.72
	OLM-SGD	70.79 ± 0.20	28.47 ± 0.55	81.89 ± 2.01	67.15 ± 0.58	81.35 ± 0.07	90.65 ± 0.31	89.58 ± 0.31	94.75 ± 0.25	61.07 ± 0.15	76.37 ± 0.49	74.21
8	OLM-SGD (without polar reconstruction)	70.65 ± 0.24	33.48 ± 1.15	90.65 ± 0.29	70.31 ± 0.10	80.53 ± 0.24	90.61 ± 0.59	91.07 ± 0.27	94.98 ± 0.34	63.49 ± 0.39	78.39 ± 0.12	76.42
	OLM-SGD	72.17 ± 0.10	34.29 ± 0.72	90.02 ± 0.06	71.80 ± 0.13	81.65 ± 0.09	91.22 ± 0.47	91.95 ± 0.08	95.16 ± 0.17	64.64 ± 0.65	78.54 ± 0.38	77.14
16	OLM-SGD (without polar reconstruction)	71.80 ± 0.11	40.98 ± 0.23	93.06 ± 0.39	75.56 ± 0.30	80.64 ± 0.11	90.93 ± 0.36	94.36 ± 0.20	95.48 ± 0.08	67.24 ± 0.63	80.32 ± 0.15	79.04
	OLM-SGD	73.50 ± 0.12	41.38 ± 0.12	93.27 ± 0.28	76.46 ± 0.37	81.92 ± 0.08	91.44 ± 0.40	95.11 ± 0.36	95.74 ± 0.18	68.75 ± 0.36	81.09 ± 0.37	79.87

Table 75: Few-shot classification on the 10 non-ImageNet datasets with the CLIP(ViT-B/16) backbone. OLM-SGD is compared with OLM-SGD (without polar reconstruction). Both methods use 70 epochs at 10^{-4} with $r = 3$. Entries are top-1 accuracy (%) as mean $\pm$ population standard deviation over seeds 1, 2, and 3. Average reports the mean across seeds after averaging the 10 datasets within each seed with equal weight. Its standard deviation is not displayed.

Shots	Method	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	CLIP	62.56	23.91	42.22	65.51	85.23	88.25	67.40	93.31	44.39	65.24	63.80
1	OLM-SGD (without polar reconstruction)	69.39 ± 0.09	28.82 ± 0.56	68.57 ± 0.26	68.18 ± 1.37	85.75 ± 0.55	92.50 ± 0.37	83.73 ± 0.23	94.00 ± 0.17	53.76 ± 0.81	74.67 ± 0.39	71.94
	OLM-SGD	70.26 ± 0.03	30.72 ± 0.49	69.74 ± 3.03	69.68 ± 1.35	85.96 ± 0.25	92.67 ± 0.29	83.31 ± 0.32	93.97 ± 0.12	53.88 ± 0.46	74.36 ± 0.29	72.45
2	OLM-SGD (without polar reconstruction)	71.24 ± 0.12	32.68 ± 1.55	78.86 ± 1.82	70.89 ± 0.47	85.82 ± 0.06	92.71 ± 0.50	89.13 ± 0.43	94.62 ± 0.30	59.87 ± 0.75	77.50 ± 0.37	75.33
	OLM-SGD	72.01 ± 0.07	33.15 ± 1.12	79.06 ± 1.66	72.60 ± 0.43	86.30 ± 0.19	92.58 ± 0.18	89.97 ± 0.38	94.47 ± 0.30	60.62 ± 0.57	77.14 ± 0.31	75.79
4	OLM-SGD (without polar reconstruction)	72.00 ± 0.18	37.20 ± 0.39	86.16 ± 0.96	74.28 ± 0.11	86.33 ± 0.15	93.50 ± 0.19	92.92 ± 0.20	95.31 ± 0.12	64.62 ± 0.92	80.08 ± 0.45	78.24
	OLM-SGD	73.39 ± 0.02	38.12 ± 0.57	84.80 ± 1.21	76.51 ± 0.18	86.83 ± 0.05	93.72 ± 0.08	93.57 ± 0.41	95.23 ± 0.20	65.03 ± 0.77	79.77 ± 0.53	78.70
8	OLM-SGD (without polar reconstruction)	73.52 ± 0.34	43.06 ± 0.10	90.60 ± 0.45	78.05 ± 0.29	86.48 ± 0.19	93.51 ± 0.28	95.07 ± 0.18	95.65 ± 0.12	68.32 ± 0.19	81.97 ± 0.47	80.62
	OLM-SGD	75.05 ± 0.18	44.85 ± 0.77	90.80 ± 0.37	79.35 ± 0.39	87.07 ± 0.14	93.76 ± 0.43	95.89 ± 0.25	95.66 ± 0.14	68.77 ± 0.58	82.29 ± 0.11	81.35
16	OLM-SGD (without polar reconstruction)	74.78 ± 0.12	50.33 ± 0.35	93.05 ± 0.26	82.17 ± 0.21	86.64 ± 0.12	93.99 ± 0.13	97.05 ± 0.20	96.27 ± 0.20	71.71 ± 0.83	84.11 ± 0.21	83.01
	OLM-SGD	76.05 ± 0.09	52.86 ± 0.57	93.24 ± 0.42	83.41 ± 0.23	87.53 ± 0.05	94.09 ± 0.14	97.62 ± 0.11	96.35 ± 0.18	72.66 ± 0.29	84.32 ± 0.10	83.81

H.6.2 Cross-Dataset Transfer of the Selected Training Epoch

Higher transfer accuracy with a smaller oracle penalty. Figure 33 and Table 76 evaluate source-selected training durations at 10^{-4}. OLM-SGD achieves 72.55% mean transfer accuracy, compared with 71.46% without polar reconstruction. Its mean target-oracle gap decreases from 1.00 to 0.37 points. It remains within one point of the target oracle on $80/90$ pairs, compared with $61/90$ for the ablation. The largest gap also decreases from 9.42 to 2.23 points. Every target has a smaller oracle gap after averaging over its nine source datasets.

The transfer gain is not explained only by a higher target-specific peak. OLM-SGD's target-oracle average is 0.47 points higher, while its transfer average is 1.10 points higher. The remaining 0.63 points come from the smaller penalty for transferring the training duration. Thus, polar reconstruction improves both attainable accuracy and retention under a source-selected stopping time.

Table 76: Epoch-transfer diagnostic for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. The target oracle averages the ten datasets' best seed-mean accuracies. Transfer accuracy and oracle gaps average the 90 off-diagonal source–target pairs. The last column counts pairs within one percentage point of the target oracle. Both methods use $r = 3$. These are test-guided diagnostics. Differences reported in the text use unrounded values.

Method	Target oracle (%)	Transfer (%)	Oracle gap (pp)	Within 1 pp
Without polar reconstruction	72.45	71.46	1.00	61/90
OLM-SGD	**72.93**	**72.55**	**0.37**	**80/90**

Different selected epochs still transfer well. OLM-SGD's selected epochs range from 57 to 292. Without polar reconstruction, they range from 3 to 241. The benefit does not require a common optimal epoch across tasks. Instead, the target trajectories retain useful accuracy over the durations selected by other datasets. The largest ablation gap occurs when Food101 selects epoch 3 for Flowers102. This transferred duration loses 9.42 points relative to the Flowers102 oracle. With OLM-SGD, no off-diagonal transfer loses more than 2.23 points.

OLM-SGD also achieves higher absolute transfer accuracy on $86/90$ pairs. Every off-diagonal entry improves over zero-shot CLIP, compared with $89/90$ without polar reconstruction. These observations support reduced dependence on the source dataset used to select the duration.

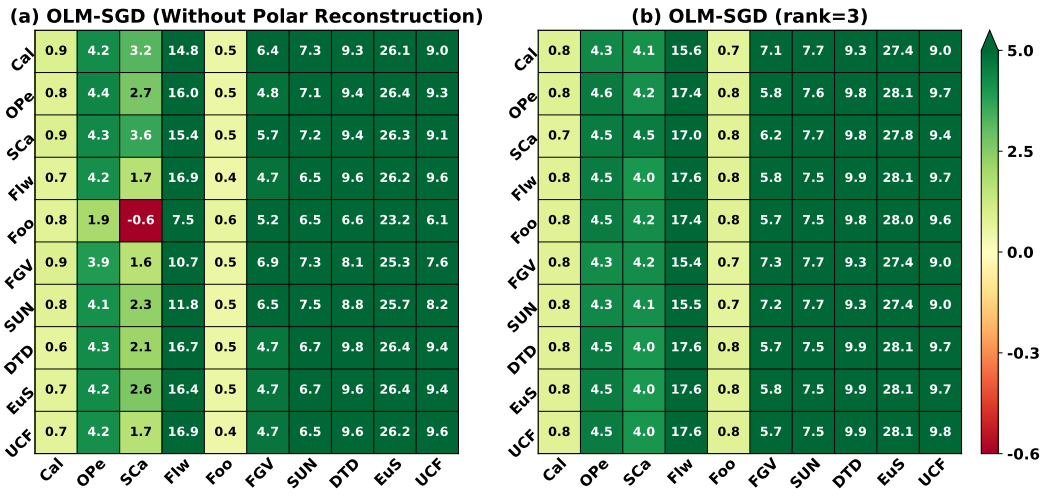

Figure 33: Cross-dataset transfer of the selected training epoch for 1-shot CLIP (ViT-B/16) at $\eta = 10^{-4}$. Both methods use $r = 3$. Each source row selects the earliest epoch in 1–300 that maximizes mean *test* accuracy across seeds 1, 2, and 3. Each target column evaluates its independently trained trajectory at that epoch. Entries give the target accuracy change relative to zero-shot CLIP, in percentage points. Green indicates improvement and red indicates degradation. Colors saturate above $+5$ points, while annotations retain the rounded numerical values. Only the duration is transferred. Diagonal entries use target-specific oracle selection. This is a test-guided diagnostic, not validation-selected early stopping.

H.6.3 Long-Horizon Dynamics and Learning-Rate Sensitivity

A broader useful learning-rate range. Figures 34 and 35 show the accuracy trajectories. Table 77 summarizes their endpoints. At 10^{-3}, OLM-SGD reaches 71.19% mean final accuracy, compared with 63.22% without polar reconstruction. It leads on nine of the ten datasets. Large gains occur on Stanford Cars (53.31% versus 31.07%), Aircraft (29.36% versus 7.22%), and EuroSAT (70.71% versus 51.51%). At this rate, the mean per-epoch population standard deviation across seeds also decreases from 3.68 to 1.20 points after averaging over datasets and epochs.

At 10^{-4}, OLM-SGD leads on all ten datasets. Its mean final accuracy is 72.68%, compared with 71.80% for direct reconstruction. Across 10^{-3}, 10^{-4}, and 10^{-5}, the range of mean final accuracy is only 1.49 points for OLM-SGD, versus 8.76 points for the ablation. Including the 10^{-2} stress test expands these ranges to 22.43 and 59.48 points. Therefore, polar reconstruction reduces sensitivity to the nominal learning rate after the rank budget and batching rule have been fixed.

Accuracy retention during prolonged training. At 10^{-4}, OLM-SGD gains 0.83 points from epoch 30 to epoch 300, whereas the ablation loses 0.18 points. From epoch 70, OLM-SGD still gains 0.22 points, while direct reconstruction loses 0.14 points. Consequently, the endpoint advantage grows beyond the fixed 70-epoch schedule.

At 10^{-5}, the final averages are nearly equal: 71.98% for OLM-SGD and 71.99% without polar reconstruction. The ablation adapts faster initially. At epoch 30, it reaches 69.96%, compared with 67.83% for OLM-SGD. Continued training adds 2.03 and 4.15 points, respectively. Thus, a sufficiently small learning rate can yield a strong endpoint without the polar stage. Polar reconstruction broadens the learning-rate range that combines useful final accuracy with strong retention.

Lower support loss does not explain the accuracy advantage. Figures 36 and 37 show support-set training losses. At 10^{-4}, direct reconstruction has lower mean final loss on every dataset, yet lower final test accuracy on every dataset. On Aircraft at 10^{-3}, its mean final loss is approximately 1.53×10^{-4}, but test accuracy is only 7.22%. OLM-SGD reaches 29.36% with a larger loss of approximately 4.27×10^{-4}. Thus, closer fitting of the scarce support data does not explain the test-accuracy advantage of polar reconstruction.

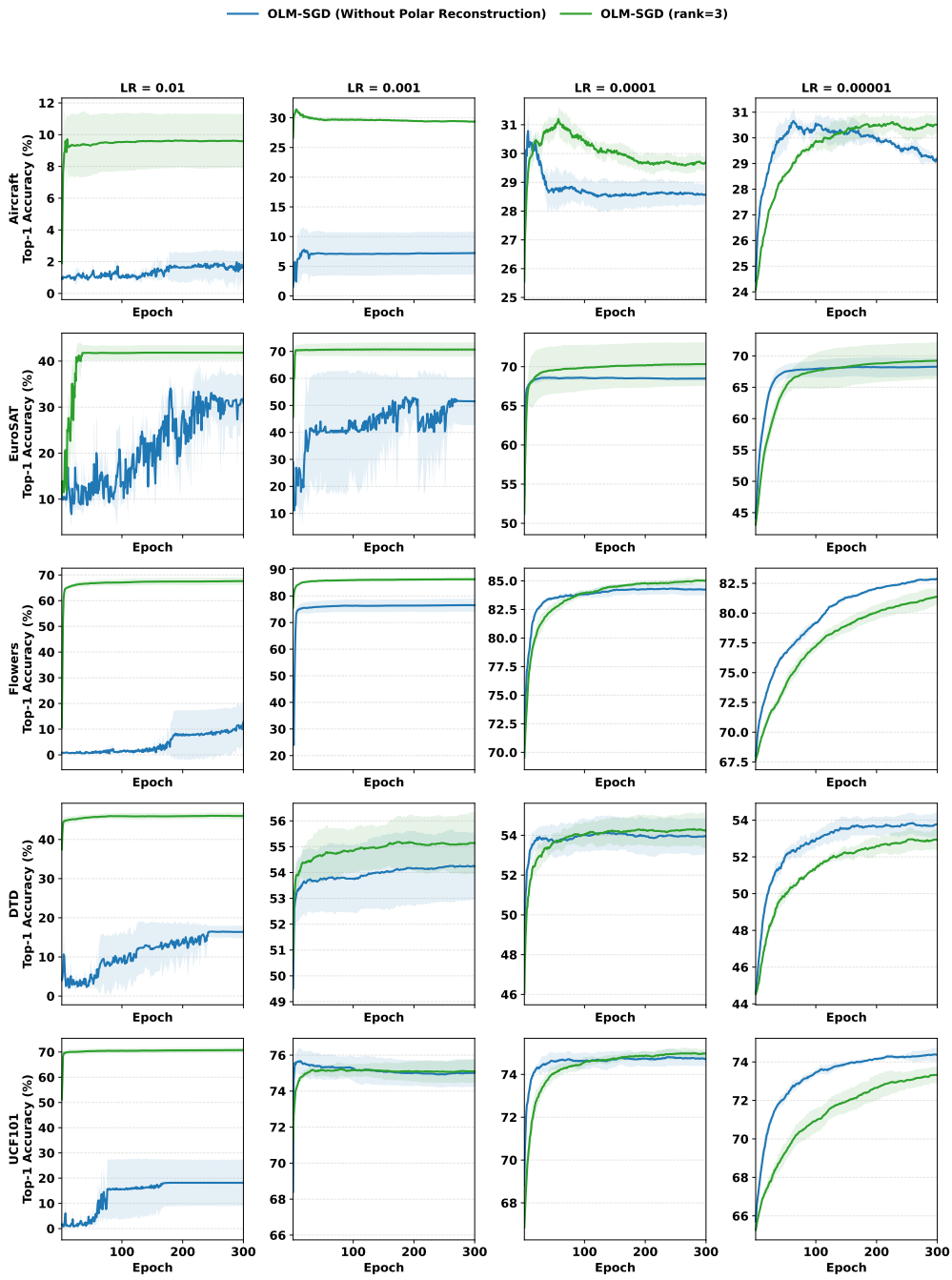

Figure 34: Test top-1 accuracy over 300 epochs of 1-shot CLIP (ViT-B/16) adaptation, with and without polar reconstruction. Both methods use $r = 3$. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across seeds 1, 2, and 3.

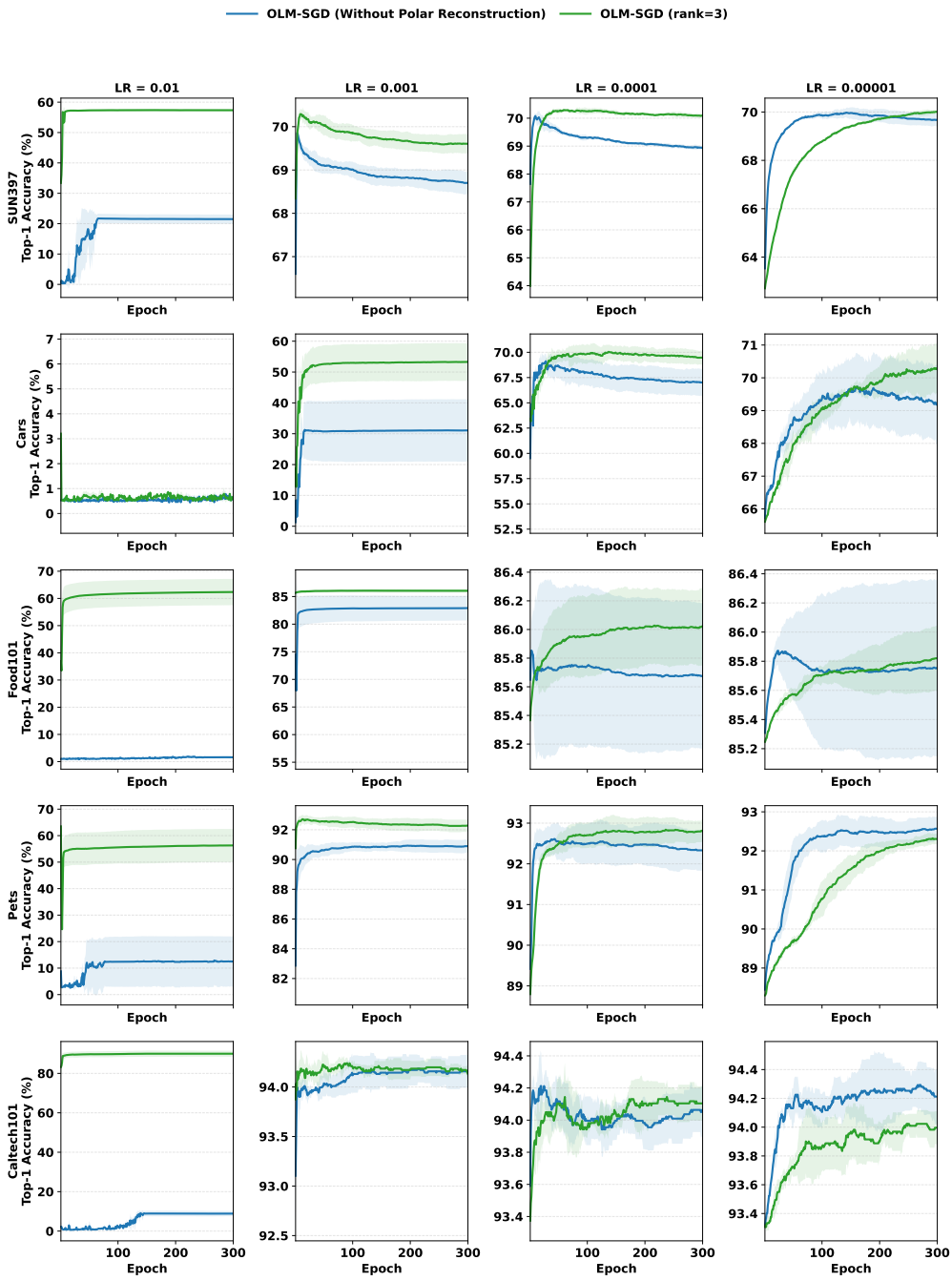

Figure 35: Test top-1 accuracy for the remaining five datasets in the polar-reconstruction comparison. Rows show SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates as Figure 34. Lines show means and shaded bands show one population standard deviation across the three seeds.

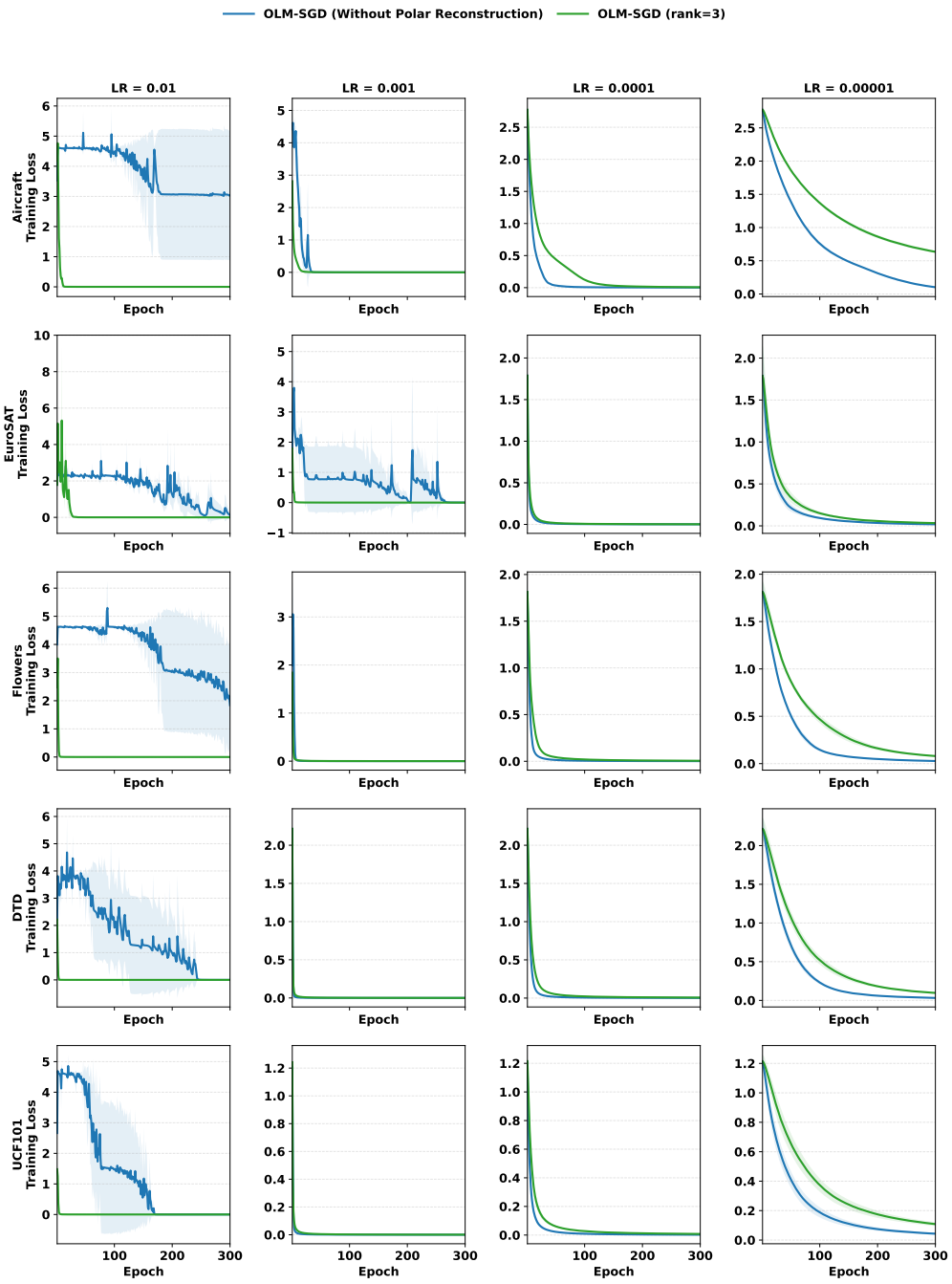

Figure 36: Training cross-entropy loss for the trajectories in Figure 34. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Lines show means and shaded bands show one population standard deviation across the three seeds. All recorded losses are finite through epoch 300.

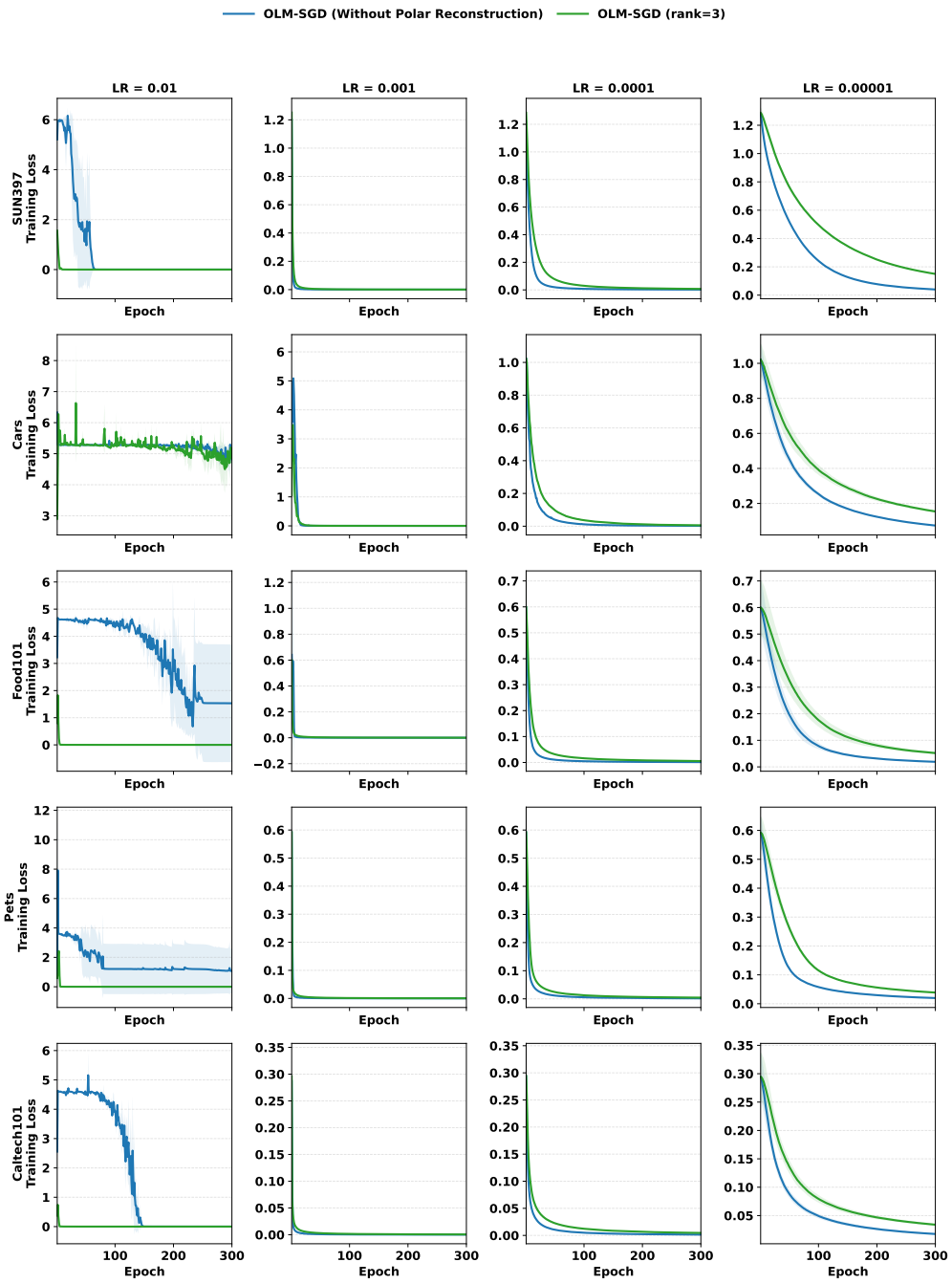

Figure 37: Training cross-entropy loss for the trajectories in Figure 35. Rows show SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns show the same four learning rates. Lines show means and shaded bands show one population standard deviation across the three seeds. All recorded losses are finite through epoch 300.

Table 77: Long-horizon summary for 1-shot CLIP (ViT-B/16) with $r = 3$. Final accuracy is evaluated at epoch 300. Best-to-final drops are computed within each seed. For both quantities, we first average the ten non-ImageNet datasets within each seed and then report mean $\pm$ population standard deviation across seeds. The last column gives the mean accuracy change from epoch 30 to epoch 300. Without polar denotes OLM-SGD without polar reconstruction. All 240 complete trajectories are included.

Learning rate	Method	Final accuracy (%)	Best-to-final (pp)	Change 30–300 (pp)
10^{-2}	Without polar	12.51 ± 2.76	1.08 ± 0.25	$+9.20$
	OLM-SGD	50.24 ± 0.48	1.33 ± 0.85	$+0.78$
10^{-3}	Without polar	63.22 ± 1.85	0.98 ± 0.19	$+1.22$
	OLM-SGD	71.19 ± 0.55	0.50 ± 0.11	$+0.15$
10^{-4}	Without polar	71.80 ± 0.11	0.78 ± 0.03	-0.18
	OLM-SGD	72.68 ± 0.32	0.31 ± 0.03	$+0.83$
10^{-5}	Without polar	71.99 ± 0.10	0.37 ± 0.03	$+2.03$
	OLM-SGD	71.98 ± 0.31	0.08 ± 0.00	$+4.15$

H.6.4 Best-to-Final Test-Accuracy Drop

Smaller degradation at the three smaller learning rates. Tables 78–81 quantify retention relative to each seed's best recorded checkpoint. At 10^{-3}, 10^{-4}, and 10^{-5}, OLM-SGD's dataset-average drops are 0.50, 0.31, and 0.08 points. Without polar reconstruction, they are 0.98, 0.78, and 0.37 points. The retention advantage therefore persists even at 10^{-5}, where the two methods have nearly identical final average accuracy.

At 10^{-4}, OLM-SGD has a smaller mean drop on every dataset. Stanford Cars decreases from 2.19 to 0.72 points. Aircraft decreases from 2.52 to 1.56 points, and SUN397 decreases from 1.20 to 0.28 points. Nine datasets remain within one point of their seed-specific peaks on average, compared with seven for the ablation. At 10^{-5}, every OLM-SGD dataset-level mean drop is below 0.30 points. Direct reconstruction still loses 1.75 points on Aircraft. These results distinguish stable late-stage adaptation from merely reaching a competitive peak.

Absolute accuracy is essential when interpreting a small drop. At 10^{-2}, the ablation has a slightly smaller average drop: 1.08 points versus 1.33 for OLM-SGD. However, its final average accuracy is only 12.51%, compared with 50.24% for OLM-SGD. OLM-SGD has higher final accuracy on every dataset at this rate. The drop metric starts at epoch 1 and excludes zero-shot CLIP. It therefore does not measure the full damage from an early loss of pretrained accuracy. For example, direct reconstruction loses only 0.85 points from its recorded peak on Caltech101, but ends at 8.86%. OLM-SGD instead retains 89.98% with a 0.07-point drop. A small best-to-final drop must therefore be read together with the absolute endpoint.

Polar reconstruction contributes beyond low-rank filtering. The two methods retain the same gradient rank budget and use the same class-unique batching rule. Their difference is how the retained contributions are combined. The squared norm of the direct average retains cross terms between the truncated gradients. Polar reconstruction instead controls the factor geometry and satisfies the retained-energy bound in (28). This operation can change both update direction and magnitude. The experiments assess their combined effect. Across the evaluated settings, the polar stage improves fixed-schedule accuracy, reduces the penalty for transferring the training duration, and broadens the useful learning-rate range. Low-rank filtering alone does not reproduce this combination of accuracy and retention.

Table 78: Best-to-final test-accuracy drop (percentage points) at learning rate 0.01 (10^{-2}), for 1-shot CLIP (ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation across seeds 1, 2, and 3. Average summarizes the three per-seed unweighted averages over all 10 datasets.

Dataset	OLM-SGD (Without Polar Reconstruction)	OLM-SGD (rank=3)
Caltech101	0.85 ± 0.38	0.07 ± 0.05
Oxford Pets	0.47 ± 0.41	9.29 ± 6.93
Stanford Cars	0.28 ± 0.10	2.83 ± 3.57
Flowers102	0.26 ± 0.36	0.04 ± 0.06
Food101	0.36 ± 0.44	0.00 ± 0.00
Aircraft	0.83 ± 0.66	0.57 ± 0.39
SUN397	0.49 ± 0.24	0.17 ± 0.17
DTD	0.99 ± 1.11	0.12 ± 0.08
EuroSAT	6.17 ± 2.28	0.17 ± 0.13
UCF101	0.09 ± 0.03	0.03 ± 0.04
Average	1.08 ± 0.25	1.33 ± 0.85

Table 79: Best-to-final test-accuracy drop (percentage points) at learning rate 0.001 (10^{-3}), for 1-shot CLIP (ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation across seeds 1, 2, and 3. Average summarizes the three per-seed unweighted averages over all 10 datasets.

Dataset	OLM-SGD (Without Polar Reconstruction)	OLM-SGD (rank=3)
Caltech101	0.12 ± 0.12	0.20 ± 0.09
Oxford Pets	0.15 ± 0.11	0.45 ± 0.15
Stanford Cars	0.56 ± 0.48	0.03 ± 0.02
Flowers102	0.07 ± 0.02	0.05 ± 0.02
Food101	0.05 ± 0.05	0.05 ± 0.02
Aircraft	2.36 ± 1.82	2.20 ± 0.39
SUN397	1.23 ± 0.22	0.74 ± 0.18
DTD	0.08 ± 0.07	0.14 ± 0.20
EuroSAT	4.33 ± 3.29	0.84 ± 0.18
UCF101	0.88 ± 0.52	0.29 ± 0.23
Average	0.98 ± 0.19	0.50 ± 0.11

Table 80: Best-to-final test-accuracy drop (percentage points) at learning rate 0.0001 (10^{-4}), for 1-shot CLIP (ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean ± population standard deviation across seeds 1, 2, and 3. Average summarizes the three per-seed unweighted averages over all 10 datasets.

Dataset	OLM-SGD (Without Polar Reconstruction)	OLM-SGD (rank=3)
Caltech101	0.27 ± 0.16	0.16 ± 0.10
Oxford Pets	0.38 ± 0.10	0.07 ± 0.05
Stanford Cars	2.19 ± 0.55	0.72 ± 0.13
Flowers102	0.09 ± 0.05	0.04 ± 0.00
Food101	0.33 ± 0.26	0.02 ± 0.02
Aircraft	2.52 ± 0.57	1.56 ± 0.22
SUN397	1.20 ± 0.10	0.28 ± 0.04
DTD	0.26 ± 0.10	0.20 ± 0.17
EuroSAT	0.45 ± 0.37	0.01 ± 0.02
UCF101	0.14 ± 0.09	0.04 ± 0.04
Average	0.78 ± 0.03	0.31 ± 0.03

Table 81: Best-to-final test-accuracy drop (percentage points) at learning rate 0.00001 (10^{-5}), for 1-shot CLIP (ViT-B/16). Lower is better. Each seed's drop is its maximum accuracy over epochs 1–300 minus its epoch-300 accuracy. Values are mean $\pm$ population standard deviation across seeds 1, 2, and 3. Average summarizes the three per-seed unweighted averages over all 10 datasets.

Dataset	OLM-SGD (Without Polar Reconstruction)	OLM-SGD (rank=3)
Caltech101	0.12 ± 0.03	0.07 ± 0.04
Oxford Pets	0.10 ± 0.11	0.02 ± 0.03
Stanford Cars	0.66 ± 0.29	0.19 ± 0.16
Flowers102	0.01 ± 0.02	0.01 ± 0.02
Food101	0.30 ± 0.36	0.06 ± 0.09
Aircraft	1.75 ± 0.33	0.29 ± 0.21
SUN397	0.35 ± 0.12	0.02 ± 0.00
DTD	0.20 ± 0.20	0.08 ± 0.06
EuroSAT	0.08 ± 0.06	0.00 ± 0.01
UCF101	0.09 ± 0.09	0.01 ± 0.01
Average	0.37 ± 0.03	0.08 ± 0.00

I Rank Sensitivity

We study the sensitivity of OLM-SGD to the retained rank $r \in \{1, 3, 5, 8\}$ under 1-shot CLIP (ViT-B/16) adaptation. For each tensor W, the effective rank is $r_W = \min\{r, m_W, n_W\}$. All other components of OLM-SGD, including class-unique mini-batching and polar reconstruction, remain unchanged. Figures 38 and 39 report the training trajectories across all 11 datasets and learning rates $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

The epoch-100 comparison uses all 11 datasets and common observed seeds across ranks. ImageNet uses seeds 1 and 2 at 10^{-4} and seeds 2 and 3 at 10^{-5}; the other dataset–rate pairs use all three seeds. Epoch-300 averages and best-to-final drops use the seven datasets with complete histories for all ranks and seeds, listed in Table 82. Incomplete runs are compared only at common observed horizons.

Table 82: Matched rank comparisons for 1-shot CLIP (ViT-B/16). A_{100} averages all 11 datasets at epoch 100 with the common seeds specified in the text. A_{300} and the best-to-final drop use only the seven datasets with complete 300-epoch histories for every rank and seed: EuroSAT, Flowers102, DTD, UCF101, Stanford Cars, Oxford Pets, and Caltech101. These two quantities report mean $\pm$ population standard deviation across seeds after averaging datasets within each seed. Datasets receive equal weight. Accuracy is in percent, and drops are in percentage points.

Learning rate	Rank	A_{100} (11 datasets)	A_{300} (7 datasets)	Drop (7 datasets)
10^{-2}	1	47.07	53.69 ± 0.95	1.94 ± 1.16
	3	47.46	53.31 ± 0.53	1.79 ± 1.22
	5	48.88	54.04 ± 0.17	1.21 ± 0.79
	8	51.24	56.28 ± 0.73	1.13 ± 0.72
10^{-3}	1	69.79	75.65 ± 0.34	0.53 ± 0.29
	3	69.76	75.27 ± 0.78	0.29 ± 0.11
	5	69.42	74.60 ± 0.78	0.19 ± 0.09
	8	69.30	74.26 ± 0.92	0.28 ± 0.09
10^{-4}	1	71.82	77.06 ± 0.47	0.20 ± 0.02
	3	72.04	77.28 ± 0.49	0.18 ± 0.01
	5	72.02	77.30 ± 0.54	0.15 ± 0.02
	8	71.94	77.12 ± 0.51	0.13 ± 0.03
10^{-5}	1	70.16	75.91 ± 0.39	0.08 ± 0.05
	3	70.45	76.20 ± 0.47	0.05 ± 0.01
	5	70.32	76.07 ± 0.49	0.06 ± 0.02
	8	70.08	75.88 ± 0.47	0.05 ± 0.03

OLM-SGD is weakly sensitive to rank in its stable learning-rate regime. The four ranks behave similarly at 10^{-4} and 10^{-5}. At 10^{-4}, their mean accuracy at epoch 100 ranges only from 71.82% to 72.04%. The largest dataset-level spread is 0.88 points. At 10^{-5}, the corresponding averages lie between 70.08% and 70.45%. The same behavior persists on the seven-dataset complete subset. At 10^{-4}, the epoch-300 mean accuracy ranges from 77.06% to 77.30%. At 10^{-5}, it ranges from 75.88% to 76.20%. Thus, increasing the retained rank does not produce a systematic accuracy gain once the learning rate is well calibrated.

Long-horizon stability is preserved across ranks. On the same seven-dataset subset at 10^{-4}, the mean best-to-final accuracy drops are 0.20, 0.18, 0.15, and 0.13 points for ranks 1, 3, 5, and 8, respectively. At 10^{-5}, all four drops remain below 0.1 points. The trajectories also show continued improvement rather than early saturation. For example, rank 3 improves from 75.54% to 81.36% on Flowers102 between epochs 70 and 300 at 10^{-5}. UCF101 improves from 70.25% to 73.33% over the same interval. These results show that the long-horizon stability of OLM-SGD is not tied to a single rank.

Rank matters more under aggressive learning rates, but no rank dominates consistently. At 10^{-3}, most datasets still show small differences across ranks. Larger deviations appear in a few datasets. For example, rank 1 performs better on Stanford Cars, whereas rank 8 performs better on ImageNet. At 10^{-2}, higher ranks improve several datasets, including Food101 and Oxford Pets, but lower ranks remain competitive on others such as DTD and UCF101. Hence, rank and learning rate play different roles. Rank controls how much spectral information is retained, whereas the learning rate has a larger effect on the overall optimization regime.

Practical rank choice. Overall, OLM-SGD operates effectively across a broad low-rank range. Rank 3 provides a strong compromise between compact spectral representations, high accuracy, and long-horizon stability. More importantly, the main empirical benefits of OLM-SGD do not depend on finely tuning r. The comparison supports the robustness of the fixed $r = 3$ setting used in the experiments.

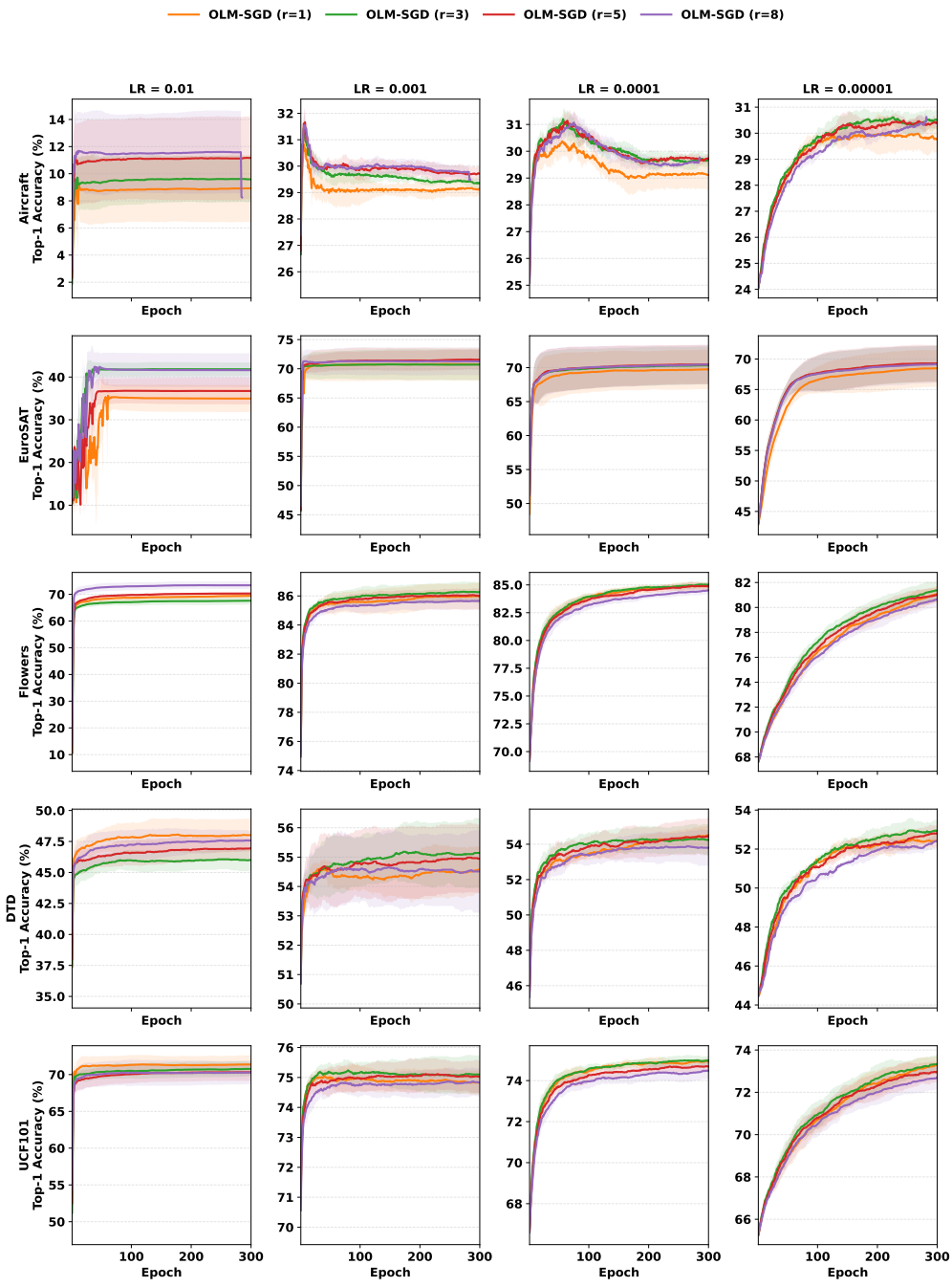

Figure 38: Rank sensitivity of OLM-SGD during 1-shot CLIP (ViT-B/16) adaptation. Rows show Aircraft, EuroSAT, Flowers102, DTD, and UCF101. Columns show learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Curves compare $r \in \{1, 3, 5, 8\}$ over up to 300 training epochs. Lines and bands show the mean and population standard deviation across available seeds among 1, 2, and 3 at each epoch. Seed counts can change for incomplete runs. Missing observations are omitted without extrapolation.

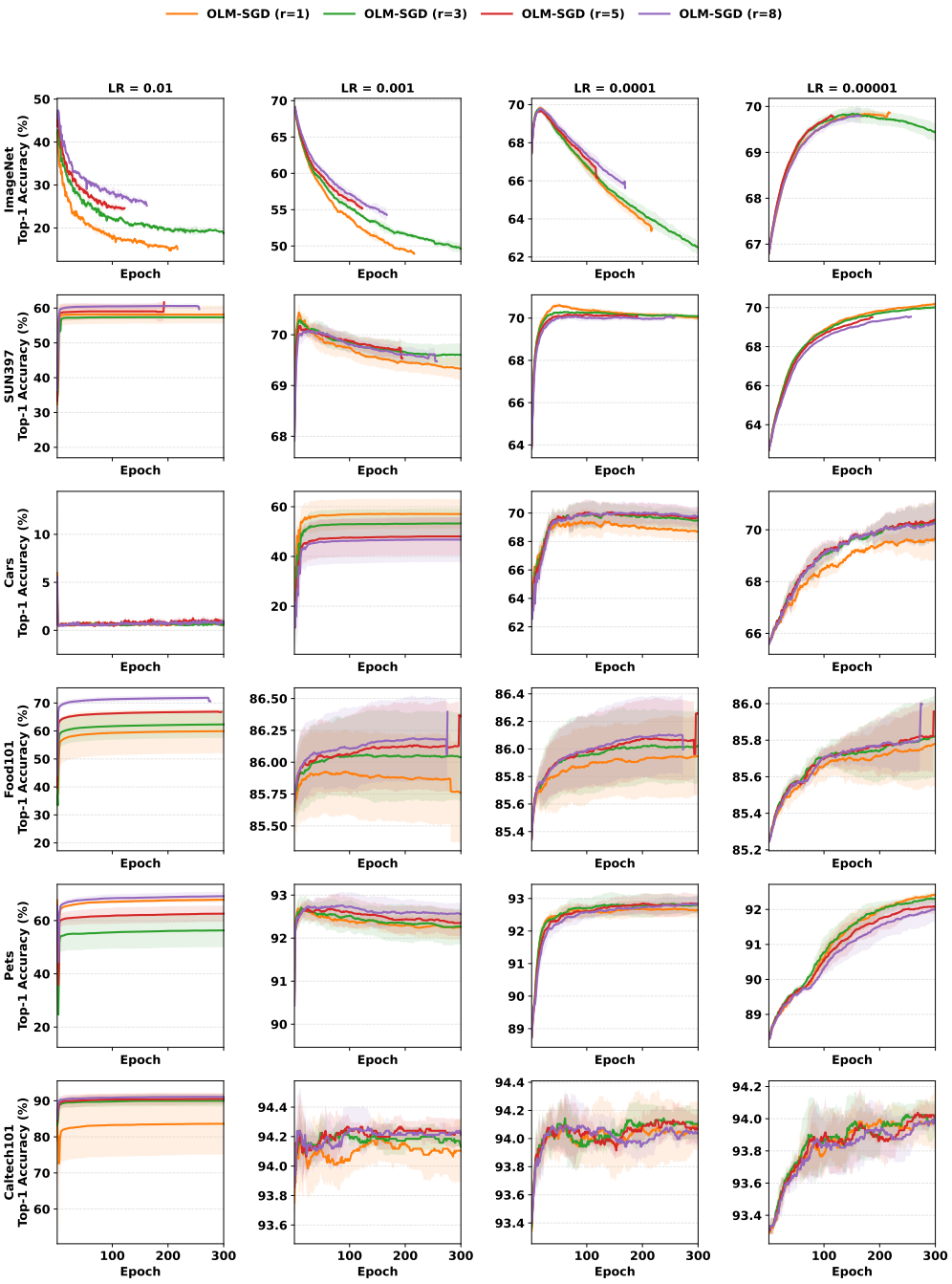

Figure 39: Rank sensitivity of OLM-SGD on ImageNet, SUN397, Stanford Cars, Food101, Oxford Pets, and Caltech101. Columns correspond to learning rates 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. Aggregation follows Figure 38. ImageNet uses the augmented extended-trajectory protocol in Appendix D. Table 82 uses common observed horizons and matched seeds.