DRIVE: Fine-tuning via Data Contribution- and Diversity-aware Weighting with Prior Regularization

Qing Liu[1] **Xinrui Chen**[1] **Yi Du**[1,2*] **Ou Wu**[1] **Weiyao Zhu**[3]

[1]Hangzhou Institute for Advanced Study,
University of Chinese Academy of Sciences, Hangzhou, China
[2]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[3]School of Mathematics, Tianjin University, Tianjin, China

{liuqing25,chenxinrui25}@mails.ucas.ac.cn

duyi@cnic.cn wuou@ucas.ac.cn weiyaozhu@tju.edu.cn

Abstract

Single-stage fine-tuning is pervasive, yet catastrophic forgetting persists when the source pre-training data is inaccessible. Data weighting affords a natural lever to modulate sample influence. Yet existing weighting methods lack a principled mechanism to explicitly arbitrate knowledge retention against new-task adaptation, and provide no unified differentiable objective to accommodate auxiliary signals such as sample diversity and sample characteristics. This is because current schemes privilege one goal over the other, and no explicit gradient-compatible encoding exists for such signals. To address this, we propose **DRIVE**, a principled and efficient weight assignment framework. It derives per-sample weights via a unified differentiable objective where three coupled signals co-evolve: Shapley additivity splits contribution into anti-forgetting and adaptation components, which are reshaped by a covariance-based diversity term and stabilized by a sample-characteristic prior. Far from a mere combination, this yields an end-to-end gradient-based scheme balancing stability against plasticity. Theoretically, we prove that DRIVE enjoys superior stability-plasticity guarantees over existing weighting schemes and enhances parameter estimation robustness. Empirically, it attains superior overall performance, faster convergence, and stronger robustness. Code at https://github.com/luciabcd/DRIVE.

1 Introduction

Fine-tuning is a key paradigm for improving downstream performance in modern deep learning [1, 2]. Recent studies show that even in a single round of fine-tuning, catastrophic forgetting can occur [3–6]. Because the original pre-training data is typically unavailable in practice, prior work has explored data-oblivious strategies in both parameter space and sample space. Parameter-space methods mitigate forgetting by constraining update magnitudes [7, 8], updating only a subset of parameters, introducing additional parameter modules [9–11], or using model merging [12, 13]. Sample-space methods mainly include sample weighting, selection, and augmentation [14–16].

Data weighting enables fine-grained control over sample influence by upweighting high-contribution samples [17–19]. In practice, most studies measure contribution by a sample's effect on new task learning, for example through weighting based on validation loss, gradient magnitude, or sample loss [14–17], whereas FLOW measures it by the sample's role in preserving pre-trained knowledge

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

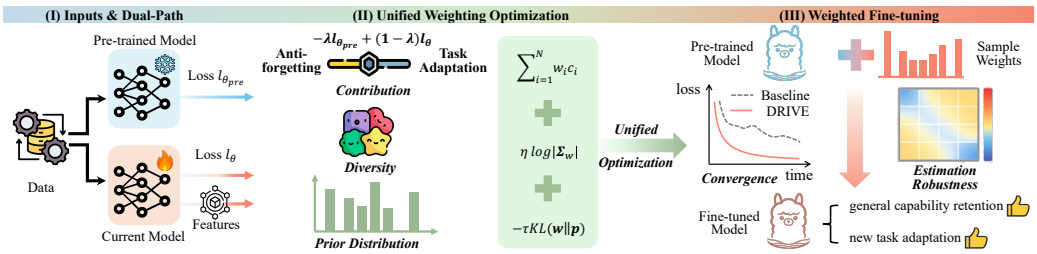


Figure 1: Overview of DRIVE's unified differentiable data weighting objective.

during fine-tuning [20]. While effective, these methods still ground sample contribution in either adaptation-oriented or retention-oriented signals, rather than explicitly balancing stability and plasticity, and thus they cannot fully characterize each sample's contribution to the fine-tuning.

Beyond sample contribution, sample diversity and sample characteristics have also been widely considered in practice. Regarding sample diversity, existing studies usually rely on clustering, DPP, or similarity coverage for selection [21–25], making it difficult to encode diversity in a continuous weighting objective. For sample characteristics, studies typically use signals such as sample length and entropy to reflect the quality or complexity of samples [26–30], and some directly treat them as proxies for sample contribution to determine sample weights [31, 32]. However, these signals are inherently task-agnostic properties with no direct correspondence to a sample's actual contribution during fine-tuning. We therefore treat them as complementary to sample contribution rather than substitutes for it. Overall, within sample weighting, existing methods usually incorporate only part of the above information into weight assignment. More importantly, because diversity is difficult to encode as a differentiable weighting term, existing sample-weighting methods lack a unified differentiable objective that jointly incorporates sample contribution, sample diversity, and sample characteristics. This not only makes it difficult to realize an end-to-end weighting scheme, limiting full data utilization, but also hinders a more principled analysis of optimization behavior and convergence.

To address the above issues, we propose **DRIVE** (Fine-tuning via **D**ata Cont**RI**bution- and Di**V**ersity-aware W**E**ighting with prior Regularization), a principled and efficient weighting framework. Rather than stacking existing signals, **DRIVE reformulates each ingredient for joint compatibility within one differentiable objective**. Sample contribution is recast through a Shapley-additivity decomposition that disentangles anti-forgetting from task adaptation and yields a tunable stability–plasticity trade-off. Diversity is made differentiable via a covariance-determinant term, while sample characteristics serve as a KL-anchored prior over weights. These reformulations enable end-to-end weighting. Theoretically, DRIVE enjoys improved stability–plasticity guarantees and enhanced parameter-estimation robustness. Empirically, across vision models and LLMs, it attains stronger overall performance. Fig. 1 illustrates the overall framework. Our key contributions are summarized as follows:

- We propose DRIVE, a novel weighting framework that jointly reformulates components rather than off-the-shelf aggregation. Shapley additivity yields an anti-forgetting and task-adaptation decomposition; a covariance-determinant term renders diversity differentiable, to our knowledge the first such term for sample weighting; and KL anchors weights to a sample-characteristic prior.
- We develop a provably convergent iterative solver for the coupled DRIVE objective. Theoretical analysis in linear/local nonlinear regimes shows contribution decomposition tightens the stability–plasticity trade-off, while diversity provably accelerates convergence and enhances robustness.
- Across vision models and LLMs, DRIVE consistently outperforms strong baselines, with better overall performance, faster convergence, and stronger robustness.

2 Related Work

Mitigating Catastrophic Forgetting in Fine-tuning. Fine-tuning is pivotal for pre-trained models [33–35], yet balancing new skills with prior capabilities remains challenging. To mitigate catastrophic forgetting, methods target parameter or sample spaces [36, 37]: regularization constrains weight drift [38–42]; optimization ensures orthogonal updates [43–45]; architectures isolate parame-

ter subsets [46–49]; and replay injects past samples [50–53]. In contrast, we introduce a data weighting perspective, optimizing the trade-off between anti-forgetting and task-adaptation.

Data Weighting in Fine-tuning. Recent data weighting strategies adopt fine-grained evaluations. Existing works optimize training dynamics via loss alignment, meta-learning, or robust dynamic weighting to mitigate noise and forgetting [16, 18, 19, 54]. Others utilize quality proxies, such as reward scores, consistency metrics, or information entropy, to guide data selection or weighting [26, 29, 30, 55–57]. Further studies isolate fine-tuning contributions through distribution shifts, trajectory analysis, or orthogonal model decomposition [58–61]. Unlike these isolated perspectives, we propose a unified weighting objective that jointly models data contribution, diversity, and sample prior.

Data Diversity in Fine-tuning. Data diversity is vital for model generalization and handling unseen semantics [62–65], supporting performance across contexts. Existing strategies rely on discrete data selection, employing clustering-based sampling to prune redundancy [23], coreset selection [22, 66], determinantal selection maximizing volume (DPP) [25], or complexity-aware filtering like DEITA [21, 24]. Others emphasize response-level diversity for alignment [67–69]. Unlike these, we integrate a covariance determinant regularizer into the loss-weighting objective.

3 Methodology

We consider a realistic deployment scenario: the single-task, data-oblivious setting of fine-tuning on a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ without access to pre-training data or continual task sequences (unlike continual learning with sequential task boundaries). We denote the parameters of pre-trained, current, and ideal models by $\boldsymbol{\theta}_{\text{pre}}$, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^*$, respectively. With model output $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ and loss $\ell(\cdot, \cdot)$, we assign a weight $w_i (\geq 0)$ to $\boldsymbol{x}_i$, satisfying $\sum_i w_i = 1$. This section infers the optimal weights via the DRIVE method, which are then employed to fine-tune the current model as shown in Fig. 1.

3.1 Data Contribution and Diversity Measures

Measuring Contribution via Shapley-inspired Scores. Data contribution quantifies the utility of individual samples for task performance. We disentangle fine-tuning contribution as balancing two distinct cooperative games over dataset $\mathcal{D}$: the anti-forgetting game v_{pre}, whose utility encodes retention of pre-trained knowledge, and the task-adaptation game v_{task}, encoding target-task performance.

Let $\phi_i(v)$ denote the Shapley value assigned to sample $\boldsymbol{x}_i$ under game v. By the **Additivity Axiom of Shapley values** [70], the contribution of sample $\boldsymbol{x}_i$ in the composite game $v_\lambda = \lambda v_{\text{pre}} + (1-\lambda) v_{\text{task}}$ with trade-off coefficient $\lambda \in [0, 1]$ must satisfy

$$\phi_i(v_\lambda) = \lambda\, \phi_i(v_{\text{pre}}) + (1 - \lambda)\, \phi_i(v_{\text{task}}). \tag{1}$$

This axiomatically justifies linear aggregation. We denote the anti-forgetting contribution by $a_i = \phi_i(v_{\text{pre}})$ and the task-adaptation contribution by $t_i = \phi_i(v_{\text{task}})$, and define the contribution as

$$c_i = C(a_i, t_i) \coloneqq \lambda\, a_i + (1 - \lambda)\, t_i. \tag{2}$$

Estimating a_i and t_i requires a per-sample contribution measure. The canonical instrument, gradient inner products against a held-out validation set [71] (see Appendix A.1), is foreclosed by our strict data-oblivious constraint, where neither pre-training data nor original validation sets are available. We therefore use a loss-based proxy computable from the model and sample alone. For cross-entropy objectives, loss measures the model's deviation from the target and provides a scalar, summary of the optimization signal carried by a sample. Thus, near-zero-loss samples typically induce limited further updates, whereas high-loss samples remain active learning signals [54]. Empirically, loss-based estimates correlate with validation-based Shapley contributions at Spearman's $\rho = 0.83$ and match gradient-based variants within 0.3% in accuracy while running $\sim 2\times$ faster (Appendices A.2 and A.3). We therefore adopt the efficient contribution estimates

$$a_i = -\ell(f(\boldsymbol{x}_i, \boldsymbol{\theta}_{\text{pre}}), \boldsymbol{y}_i), \qquad t_i = \ell(f(\boldsymbol{x}_i, \boldsymbol{\theta}), \boldsymbol{y}_i). \tag{3}$$

Measuring Diversity via the Covariance Determinant. Diversity captures set-level characteristics across samples and is critical for reducing redundancy and improving downstream generalization [21–24]. The most natural choice, clustering-based selection, is discrete and non-differentiable and thus incompatible with our continuous weight optimization. Yet a differentiable, principled set-level alternative is nontrivial to construct: combinatorial coverage terms are non-smooth, pairwise-repulsion

surrogates scale quadratically and conflate global coverage with local separation, and even DPPs encode set-level volume only over discrete subsets. We therefore lift this set-level volume principle to continuous data weighting via D-optimal design [72, 73], to our knowledge the first such adaptation.

Under weights $\boldsymbol{w} = \{w_i\}$, we define $\boldsymbol{\Sigma}_w = \sum_{i=1}^{N} w_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top$ with $\bar{\boldsymbol{x}}_i$ the fixed, unweightedly centered feature vector from the current model, and set $\mathcal{D}(\boldsymbol{w}) := \log|\boldsymbol{\Sigma}_w|$, which tracks the evolving representation rather than frozen pre-trained features. The objective is smooth in $\boldsymbol{w}$ with closed-form gradient $\frac{\partial}{\partial w_i} \log|\boldsymbol{\Sigma}_w| = \bar{\boldsymbol{x}}_i^\top \boldsymbol{\Sigma}_w^{-1} \bar{\boldsymbol{x}}_i$, the statistical leverage of sample i, so upweighting yields a large marginal diversity gain precisely when its features are poorly explained by the rest. This makes $\mathcal{D}(\boldsymbol{w})$ a principled, differentiable plug-in for the weighting objective introduced next.

3.2 Data Weight Optimization Objective via Prior Regularization

Beyond contribution and diversity, samples possess intrinsic characteristics such as quality and complexity that are properties of the data itself and largely task-agnostic. Folding such information into c_i would conflate these sample-intrinsic properties with the task-specific contribution it is meant to capture. We instead encode them as a prior distribution $\boldsymbol{p} = \{p_i\}$ over samples, which serves as a KL-anchored reference that regularizes the weight distribution toward sample-intrinsic preferences and jointly determines sample weights together with contribution and diversity.

Given the above, data weights are obtained by solving the following optimization objective

$$\{w_i^*\} = \arg\max_{\boldsymbol{w}} J(\boldsymbol{w}), \quad J(\boldsymbol{w}) = \sum_{i=1}^{N} w_i c_i + \eta \log|\boldsymbol{\Sigma}_w| - \tau KL(\boldsymbol{w}\|\boldsymbol{p}) \quad \text{s.t.} \sum_{i=1}^{N} w_i = 1,\ w_i \geq 0. \tag{4}$$

Here, the first term of $J(\boldsymbol{w})$ captures sample contribution, the second promotes diversity through feature-volume expansion with strength η, and the third incorporates a prior reference distribution $\boldsymbol{p}$ with strength τ. The stationarity condition of Eq. (4) admits a closed-form expression for w_i, but it cannot be evaluated directly because $\boldsymbol{\Sigma}_w$ on its right-hand side is itself a function of $\boldsymbol{w}$. We therefore solve this fixed-point equation by iteration, which we observe to converge within a few steps. Starting from the maximum-entropy uniform $\boldsymbol{w}^{(0)}$ with induced covariance $\boldsymbol{\Sigma}_w^{(0)}$, each iteration $t \geq 1$ uses the previous covariance $\boldsymbol{\Sigma}_w^{(t-1)}$ in place of $\boldsymbol{\Sigma}_w$ to update the weights as

$$w_i^{(t)} \propto p_i \exp\left(\frac{c_i + \eta\, \bar{\boldsymbol{x}}_i^\top (\boldsymbol{\Sigma}_w^{(t-1)})^{-1} \bar{\boldsymbol{x}}_i}{\tau}\right). \tag{5}$$

The updated weights are normalized to satisfy the simplex constraint, and the procedure iterates until convergence or the maximum iteration count is reached. This update admits the interpretation of entropic mirror ascent under generalized KL regularization, with the KL term serving as the proximal regularizer. Moreover, for $\tau > 0$, $J(\boldsymbol{w})$ is strictly concave on the simplex, so Eq. (4) admits a unique global maximizer; the fixed-point iteration converges linearly under the standard contraction condition. Detailed derivations and proofs are provided in Sec. 4.2 and Appendix B.

In this study, we adopt a uniform prior (i.e., $p_i \equiv \frac{1}{N}$) as the default non-informative choice in our main experiments, since it introduces minimal assumptions. When other signals are available, however, the framework naturally allows more informative priors; recent studies suggest that sample length or entropy can serve as proxies for sample quality in sample selection [26, 29, 30], and our ablation in Sec. 5.3 further supports the effectiveness of entropy-informed priors. Under a uniform prior, the objective admits several special cases (Table 1).

Table 1: Special cases of DRIVE under a uniform prior

Case	Weight form	Meaning
$\eta = 0$	$w_i^* \propto \exp[(-\lambda l_i^{\theta_{\text{pre}}} + (1-\lambda) l_i^{\theta})/\tau]$	Without diversity
- $\lambda = 1$	$w_i^* \propto \exp(-l_i^{\theta_{\text{pre}}}/\tau)$	Anti-forgetting [20]
- $\lambda = 0$	$w_i^* \propto \exp(l_i^{\theta}/\tau)$	Task adaptation [16]
- $\lambda = 0.5$	$\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{pre}} \Rightarrow w_i^* \propto 1/N$	Standard FT (init.)
$\eta > 0$	$w_i^{(t)} \propto \exp[(c_i + \eta\, \bar{\boldsymbol{x}}_i^\top (\boldsymbol{\Sigma}_w^{(t-1)})^{-1} \bar{\boldsymbol{x}}_i)/\tau]$	Full DRIVE

We primarily anneal λ via a linear decay ($1 \to 0$) over fine-tuning, with a task-specific constant as alternative. The decay enacts a curriculum: large λ early preserves pretrained capability, while $\lambda \to 0$ late prioritizes task adaptation. This mirrors the recall-then-learn principle of [74], which anneals a pretraining-recall loss against the task loss, thereby reconciling early stabilization with late adaptation without per-task tuning (Table 8).

Fig. 2 shows the impacts of the uniform prior and diversity terms, as well as their synergistic effect in a 1D scenario. In Fig. 2(a), the prior term enforces equal sample weights, causing the model to over-concentrate on dense regions dominated by redundant samples. Conversely, maximizing the diversity term alone (Fig. 2(b)) biases weights toward peripheral samples at the expense of the majority of data. DRIVE (Fig. 2(c)) reconciles these tendencies by integrating both; it up-weights peripheral samples to promote diversity while anchoring to the uniform prior to ensure coverage. We prove in Sec. 4 that incorporating diversity accelerates convergence and enhances estimation robustness. Conventional diversity strategies, such as clustering, are typically discrete and hard to incorporate into unified objectives. Our determinant-based approach offers a differentiable alternative that empirically outperforms the heuristic clustering method DEFT-UCS [22] in Sec. 5.2.

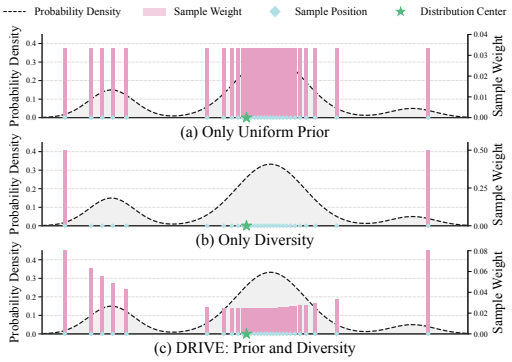


Figure 2: Impact of Uniform Prior and Determinant Regularization on Weight Distribution.

3.3 Efficient Data Weighting for Fine-tuning

While the iterative solver for Eq. (4) is theoretically convergent, using it in fine-tuning faces a key challenge: inverting $\mathbf{\Sigma}_w$ costs $O(d^3)$, prohibitive for high-dimensional $\boldsymbol{x} \in \mathbb{R}^d$.

To bound the inversion cost, we sketch the centered feature $\tilde{\boldsymbol{x}}_i$, extracted from the current model's penultimate layer (Vision) [75] or final-block layer (LLM) [76], through a Gaussian random matrix $\mathbf{R} \in \mathbb{R}^{d\times k}$ with $k \ll d$, giving $\tilde{\boldsymbol{x}}_i = \mathbf{R}^\top \tilde{\boldsymbol{x}}_i \in \mathbb{R}^k$.

Algorithm 1 DRIVE (Given step-wise inputs c_i and $\tilde{\boldsymbol{x}}_i$.)

Require: $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^S$, c_i, $\tilde{\boldsymbol{x}}_i$, k, τ, η, ϵ, $\tilde{w}_i^{(0)} \leftarrow 1/S$.
Ensure: Optimal weights $\boldsymbol{w}^*$.

```
Generate a Gaussian random matrix R ∈ R^{d×k}.
Project features: x̃_i ← R^⊤ x̃_i for all i = 1, ..., S.
for t = 1 to T_max do
    Σ_w^{(t-1)} ← Σ_{i=1}^S w̃_i^{(t-1)} x̃_i x̃_i^⊤ + εI
    Compute w_i^{(t)} via Eq. (5) for all i
    w̃_i^{(t)} ← w_i^{(t)} / Σ_{j=1}^S w_j^{(t)}
    If converged, then break
end for
return w* ← w̃^{(t)}
```

The Johnson–Lindenstrauss lemma suggests that the Gaussian sketch approximately preserves inner products and pairwise distances on the dominant low-rank local subspace (Appendix B.5), while reducing inversion to $O(k^3)$. $\mathbf{R}$ is generated once and shared across all steps, and ensembling projections yields negligible gains (Appendix F.3). Since a single mini-batch $\mathcal{B}$ may not contain enough samples for $\mathbf{\Sigma}_w$ to be invertible, we augment it with a FIFO buffer of the 128 most recently seen samples, forming a local support set of size S on which both $\mathbf{\Sigma}_w$ and the centering mean are estimated. We set $k = 128$ to match the buffer size, aligning the sketch dimension with the local support scale and avoiding overly aggressive compression. Sensitivity to k is reported in Appendix F.5. For numerical stability, we add a small ridge $\epsilon\mathbf{I}$ to $\mathbf{\Sigma}_w$ before inversion. Algorithm 1 summarizes the per-step procedure. At step s, we compute c_i and $\tilde{\boldsymbol{x}}_i$ for the current batch and reuse cached FIFO values to form the local support set, then run the projected fixed-point iteration to obtain $\boldsymbol{w}^*$. We restrict and renormalize $\boldsymbol{w}^*$ within $\mathcal{B}$, form the weighted loss, and update $\boldsymbol{\theta}^{(s)}$ to $\boldsymbol{\theta}^{(s+1)}$.

4 Theoretical Analysis

Following FLOW [20] and related analyses of weighted fine-tuning, we study DRIVE in both linear and nonlinear regimes. The linear case serves as a standard analytical abstraction, widely adopted in prior work [77, 78]. To better reflect deep fine-tuning in practice, we further extend the analysis to the nonlinear regime under local smoothness and local linearization assumptions. In addition, we analyze the iterative solver for Eq. (4) and establish the linear convergence of the resulting fixed-point iteration. The empirical results in Secs. 5.2 and 5.5 are consistent with these theoretical insights.

Table 2: Performance comparison of models fine-tuned on vision benchmarks (Accuracy in %). Best in **bold**, second in underlined.

Model	Method	CIFAR10			CIFAR100			Flowers102		
		IN-1K Acc.	Target Acc.	Avg Acc.	IN-1K Acc.	Target Acc.	Avg Acc.	IN-1K Acc.	Target Acc.	Avg Acc.
ResNet-18	Pre-trained	**69.76**$_{\pm 0.00}$	–	–	**69.76**$_{\pm 0.00}$	–	–	**69.76**$_{\pm 0.00}$	–	–
	Standard FT	34.53$_{\pm 0.03}$	**96.13**$_{\pm 0.06}$	65.33$_{\pm 0.04}$	3.28 $_{\pm 0.06}$	**84.12**$_{\pm 0.04}$	43.70$_{\pm 0.05}$	5.28$_{\pm 0.01}$	92.00$_{\pm 0.03}$	48.64$_{\pm 0.02}$
	Linear Probe	**69.76**$_{\pm 0.00}$	81.24$_{\pm 0.12}$	75.50$_{\pm 0.06}$	**69.76**$_{\pm 0.00}$	59.01$_{\pm 0.02}$	64.39$_{\pm 0.01}$	**69.76**$_{\pm 0.00}$	86.86$_{\pm 0.02}$	78.31$_{\pm 0.01}$
	WiSE-FT	57.04$_{\pm 0.07}$	87.91$_{\pm 0.03}$	72.48$_{\pm 0.03}$	49.31$_{\pm 0.04}$	62.18$_{\pm 0.01}$	55.75$_{\pm 0.02}$	49.60$_{\pm 0.01}$	75.69$_{\pm 0.05}$	62.65$_{\pm 0.02}$
	ℓ_2-Reg.	37.86$_{\pm 0.05}$	95.53$_{\pm 0.04}$	66.70$_{\pm 0.05}$	29.86$_{\pm 0.04}$	81.82$_{\pm 0.07}$	55.84$_{\pm 0.06}$	19.34$_{\pm 0.06}$	**92.11**$_{\pm 0.06}$	55.73$_{\pm 0.04}$
	DEFT-UCS	47.56$_{\pm 0.05}$	87.26$_{\pm 0.06}$	67.41$_{\pm 0.06}$	49.63$_{\pm 0.07}$	79.63$_{\pm 0.05}$	64.63$_{\pm 0.06}$	62.45$_{\pm 0.04}$	87.94$_{\pm 0.07}$	75.20$_{\pm 0.06}$
	FLOW	68.83$_{\pm 0.02}$	88.44$_{\pm 0.05}$	78.64$_{\pm 0.04}$	53.05$_{\pm 0.07}$	77.07$_{\pm 0.02}$	65.06$_{\pm 0.02}$	69.47$_{\pm 0.06}$	87.64$_{\pm 0.02}$	78.56$_{\pm 0.04}$
	DRIVE (Ours)	67.17$_{\pm 0.03}$	92.38$_{\pm 0.04}$	**79.78**$_{\pm 0.03}$	54.66$_{\pm 0.01}$	80.39$_{\pm 0.01}$	**67.53**$_{\pm 0.01}$	68.20$_{\pm 0.02}$	90.65$_{\pm 0.05}$	**79.43**$_{\pm 0.03}$
ResNet-50	Pre-trained	**80.10**$_{\pm 0.00}$	–	–	**80.10**$_{\pm 0.00}$	–	–	**80.10**$_{\pm 0.00}$	–	–
	Standard FT	29.99$_{\pm 0.04}$	**97.19**$_{\pm 0.03}$	63.59$_{\pm 0.03}$	2.50$_{\pm 0.06}$	**86.10**$_{\pm 0.07}$	44.30$_{\pm 0.05}$	6.70$_{\pm 0.07}$	**92.70**$_{\pm 0.06}$	49.70$_{\pm 0.05}$
	Linear Probe	**80.10**$_{\pm 0.00}$	82.46$_{\pm 0.03}$	81.28$_{\pm 0.02}$	**80.10**$_{\pm 0.00}$	61.11$_{\pm 0.05}$	70.61$_{\pm 0.02}$	**80.10**$_{\pm 0.00}$	85.84$_{\pm 0.04}$	82.97$_{\pm 0.02}$
	WiSE-FT	63.17$_{\pm 0.04}$	91.71$_{\pm 0.02}$	77.44$_{\pm 0.02}$	55.42$_{\pm 0.03}$	71.33$_{\pm 0.05}$	63.38$_{\pm 0.03}$	58.07$_{\pm 0.03}$	85.82$_{\pm 0.07}$	71.95$_{\pm 0.03}$
	ℓ_2-Reg.	33.98$_{\pm 0.05}$	97.00$_{\pm 0.03}$	65.49$_{\pm 0.02}$	47.16$_{\pm 0.03}$	85.77$_{\pm 0.01}$	66.47$_{\pm 0.02}$	62.85$_{\pm 0.06}$	91.67$_{\pm 0.03}$	77.26$_{\pm 0.05}$
	DEFT-UCS	61.73$_{\pm 0.07}$	90.92$_{\pm 0.03}$	76.33$_{\pm 0.05}$	57.92$_{\pm 0.06}$	82.11$_{\pm 0.08}$	70.02$_{\pm 0.07}$	70.42$_{\pm 0.05}$	89.77$_{\pm 0.09}$	80.10$_{\pm 0.07}$
	FLOW	75.15$_{\pm 0.04}$	89.18$_{\pm 0.02}$	82.17$_{\pm 0.02}$	60.85$_{\pm 0.05}$	80.74$_{\pm 0.05}$	70.80$_{\pm 0.03}$	72.45$_{\pm 0.01}$	88.93$_{\pm 0.07}$	80.69$_{\pm 0.04}$
	DRIVE (Ours)	73.89$_{\pm 0.03}$	94.45$_{\pm 0.03}$	**84.17**$_{\pm 0.03}$	61.64$_{\pm 0.05}$	83.03$_{\pm 0.06}$	**72.34**$_{\pm 0.02}$	75.02$_{\pm 0.01}$	92.67$_{\pm 0.04}$	**83.85**$_{\pm 0.01}$

4.1 Linear and Local Nonlinear Regimes

Linear regime. We first consider a linear regression abstraction, following prior analyses of FLOW [20]. Under sample weights $w(\boldsymbol{x})$, the training dynamics are governed by the weighted covariance $\boldsymbol{\Sigma}_w = \mathbb{E}[w(\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^\top]$, and convergence along direction $\boldsymbol{u}$ is controlled by $\mu_{\boldsymbol{u}} = \boldsymbol{u}^\top \boldsymbol{\Sigma}_w \boldsymbol{u}$. For directions where pre-training and fine-tuning conflict, stability-oriented weighting such as FLOW can suppress the spectrum and stall adaptation toward the ideal target $\boldsymbol{\theta}^*$. In contrast, DRIVE uses $c_i = \lambda a_i + (1-\lambda) t_i$, so decreasing λ reallocates weight to task-relevant conflicting samples, restores learning in those directions, and reduces the fitting gap to $\boldsymbol{\theta}^*$ when the local approximation error is controlled. Moreover, maximizing $\log|\boldsymbol{\Sigma}_w|$ is equivalent to D-optimal design under the simplex constraint: it enlarges the weighted feature volume, reduces generalized variance, and improves the condition number of $\boldsymbol{\Sigma}_w$, thereby accelerating convergence and enhancing estimator stability. Therefore, in the linear regime, DRIVE improves both adaptation toward $\boldsymbol{\theta}^*$ and optimization robustness.

Local nonlinear regime. We consider deep fine-tuning under local smoothness around $\boldsymbol{\theta}_{\text{pre}}$. In this setting, the weighted Hessian decomposes into a Gauss–Newton component, a non-Gauss–Newton residual, and a local nonlinear remainder. The principal Gauss–Newton block is represented by the weighted matrix $\boldsymbol{G}_w(\boldsymbol{\theta})$. Stability-oriented weighting can assign insufficient mass to task-relevant conflicting directions, slowing movement toward $\boldsymbol{\theta}^*$. Shifting λ from anti-forgetting toward task adaptation reallocates weight to these directions and restores the spectrum of $\boldsymbol{G}_w(\boldsymbol{\theta})$, reducing the local residual gradient toward $\boldsymbol{\theta}^*$ when the nonlinear remainder is controlled. Meanwhile, the diversity term $\log|\boldsymbol{\Sigma}_w|$, defined on the centered feature covariance, emphasizes under-covered feature directions and discourages redundancy. This promotes a better-covered representation geometry, which supports better-conditioned local optimization through the weighted Gauss–Newton block.

4.2 Convergence Analysis of the Iterative Solver for Eq. (4)

With the ridge-regularized covariance used in the solver, $J(\boldsymbol{w})$ is τ-strongly concave on Δ_{N-1} for $\tau > 0$ and therefore Eq. (4) admits a unique maximizer $\boldsymbol{w}^*$. Let $\{\boldsymbol{w}^{(t)}\}_{t\geq 0}$ be the sequence generated by Eq. (5). If the diversity-gradient term $\boldsymbol{h}(\boldsymbol{w})$ is L_h-Lipschitz and $\eta L_h < \tau$, then the fixed-point map is contractive, so there exists $\xi = \eta L_h/\tau < 1$ such that $\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_1 \leq \xi^t \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|_1$. Therefore, the iterate after T steps satisfies $\|\boldsymbol{w}^{(T)} - \boldsymbol{w}^*\|_1 = O(\xi^T)$. Hence, under the stated contraction condition, the proposed solver returns a near-optimal weight vector after finitely many iterations. Proofs are deferred to Appendix B and Appendix C.

5 Experiments

Our comprehensive empirical analysis is structured to address five core research questions: ❶ Does DRIVE outperform SOTAs in balancing task adaptation and anti-forgetting capabilities? ❷ Do indi-

Table 3: Performance comparison of models fine-tuned on the MetaMathQA dataset (Accuracy in %). Best in **bold**, second in underlined.

Model	Method	General Capability Score (%)				Target Acc. (%)	Overall (%)
		CS	MMLU	MBPP	Avg.	GSM8K	Avg.
Gemma-2-2B	Pre-trained	$57.23_{\pm 0.00}$	$53.14_{\pm 0.00}$	$28.60_{\pm 0.00}$	$46.32_{\pm 0.00}$	$23.65_{\pm 0.00}$	$34.99_{\pm 0.00}$
	Standard FT	$54.89_{\pm 1.01}$	$49.26_{\pm 0.22}$	$23.40_{\pm 2.29}$	$42.52_{\pm 0.82}$	$\mathbf{63.81}_{\pm 0.67}$	$53.17_{\pm 0.73}$
	WiSE-FT	$55.28_{\pm 0.62}$	$50.42_{\pm 0.50}$	$26.80_{\pm 1.41}$	$44.17_{\pm 0.65}$	$54.00_{\pm 0.51}$	$49.09_{\pm 0.44}$
	ℓ_2-Reg.	$\mathbf{58.76}_{\pm 0.97}$	$\mathbf{53.92}_{\pm 0.15}$	$30.72_{\pm 1.08}$	$\mathbf{47.80}_{\pm 0.48}$	$58.50_{\pm 0.44}$	$53.15_{\pm 0.40}$
	LoRA	$55.58_{\pm 0.47}$	$50.34_{\pm 0.39}$	$27.20_{\pm 1.55}$	$44.37_{\pm 0.67}$	$61.50_{\pm 0.21}$	$52.94_{\pm 0.36}$
	DEFT-UCS	$52.05_{\pm 0.75}$	$44.55_{\pm 0.23}$	$22.80_{\pm 1.44}$	$39.80_{\pm 0.67}$	$56.63_{\pm 0.44}$	$48.22_{\pm 0.42}$
	FLOW	$58.00_{\pm 2.35}$	$53.50_{\pm 0.32}$	$31.00_{\pm 2.38}$	$47.50_{\pm 1.46}$	$60.20_{\pm 0.49}$	$53.85_{\pm 0.82}$
	TALR	$57.98_{\pm 1.87}$	$53.58_{\pm 0.26}$	$31.04_{\pm 1.22}$	$47.53_{\pm 0.67}$	$60.20_{\pm 0.48}$	$53.87_{\pm 0.35}$
	DRIVE (Ours)	$57.92_{\pm 1.46}$	$53.59_{\pm 0.14}$	$\mathbf{31.40}_{\pm 1.90}$	$47.64_{\pm 1.05}$	$63.46_{\pm 0.43}$	$\mathbf{55.55}_{\pm 0.45}$
Llama-3.2-3B	Pre-trained	$54.62_{\pm 0.00}$	$56.54_{\pm 0.00}$	$38.60_{\pm 0.00}$	$49.92_{\pm 0.00}$	$27.75_{\pm 0.00}$	$38.84_{\pm 0.00}$
	Standard FT	$52.00_{\pm 1.03}$	$52.50_{\pm 0.43}$	$33.52_{\pm 2.20}$	$46.01_{\pm 1.13}$	$\mathbf{65.81}_{\pm 1.05}$	$55.91_{\pm 0.08}$
	WiSE-FT	$54.57_{\pm 1.40}$	$52.01_{\pm 0.94}$	$33.80_{\pm 1.58}$	$46.79_{\pm 0.70}$	$56.01_{\pm 0.41}$	$51.40_{\pm 0.49}$
	ℓ_2-Reg.	$\mathbf{56.01}_{\pm 0.60}$	$51.99_{\pm 0.45}$	$40.52_{\pm 1.59}$	$49.51_{\pm 0.62}$	$60.00_{\pm 0.89}$	$54.76_{\pm 0.44}$
	LoRA	$54.17_{\pm 1.16}$	$53.33_{\pm 0.22}$	$35.92_{\pm 1.41}$	$47.81_{\pm 0.48}$	$63.50_{\pm 0.48}$	$55.66_{\pm 0.16}$
	DEFT-UCS	$51.28_{\pm 1.55}$	$46.03_{\pm 0.38}$	$23.80_{\pm 1.58}$	$40.37_{\pm 0.83}$	$60.38_{\pm 0.58}$	$50.38_{\pm 0.49}$
	FLOW	$56.00_{\pm 1.19}$	$\mathbf{56.50}_{\pm 0.33}$	$41.12_{\pm 2.33}$	$51.21_{\pm 1.06}$	$61.20_{\pm 0.48}$	$56.21_{\pm 0.48}$
	TALR	$55.98_{\pm 2.35}$	$56.43_{\pm 0.58}$	$\mathbf{41.96}_{\pm 1.93}$	$\mathbf{51.46}_{\pm 1.46}$	$61.05_{\pm 0.38}$	$56.26_{\pm 0.61}$
	DRIVE (Ours)	$55.92_{\pm 2.68}$	$55.71_{\pm 0.28}$	$41.00_{\pm 1.08}$	$50.88_{\pm 1.05}$	$64.50_{\pm 0.53}$	$\mathbf{57.69}_{\pm 0.36}$
Llama-3.1-8B	Pre-trained	$61.59_{\pm 0.00}$	$65.32_{\pm 0.00}$	$47.40_{\pm 0.00}$	$58.10_{\pm 0.00}$	$50.95_{\pm 0.00}$	$54.53_{\pm 0.00}$
	Standard FT	$56.50_{\pm 1.03}$	$58.50_{\pm 0.37}$	$33.52_{\pm 1.14}$	$49.51_{\pm 0.33}$	$76.91_{\pm 0.56}$	$63.21_{\pm 0.22}$
	WiSE-FT	$57.46_{\pm 0.58}$	$62.37_{\pm 0.58}$	$40.60_{\pm 1.39}$	$53.48_{\pm 0.80}$	$72.89_{\pm 0.66}$	$63.19_{\pm 0.36}$
	ℓ_2-Reg.	$60.09_{\pm 1.27}$	$62.06_{\pm 0.56}$	$44.60_{\pm 2.32}$	$55.58_{\pm 1.14}$	$73.21_{\pm 0.78}$	$64.40_{\pm 0.83}$
	LoRA	$60.90_{\pm 1.20}$	$63.40_{\pm 0.44}$	$46.40_{\pm 0.51}$	$56.90_{\pm 0.59}$	$74.50_{\pm 0.42}$	$65.70_{\pm 0.49}$
	DEFT-UCS	$60.78_{\pm 0.41}$	$63.35_{\pm 0.19}$	$38.20_{\pm 2.37}$	$54.11_{\pm 0.87}$	$74.37_{\pm 0.43}$	$64.24_{\pm 0.43}$
	FLOW	$60.00_{\pm 0.65}$	$63.00_{\pm 0.44}$	$\mathbf{48.00}_{\pm 1.70}$	$57.00_{\pm 0.70}$	$73.90_{\pm 0.16}$	$65.45_{\pm 0.40}$
	TALR	$60.82_{\pm 0.77}$	$63.36_{\pm 0.26}$	$47.48_{\pm 2.16}$	$57.22_{\pm 0.81}$	$73.84_{\pm 0.41}$	$65.53_{\pm 0.51}$
	DRIVE (Ours)	$\mathbf{60.95}_{\pm 0.54}$	$\mathbf{63.98}_{\pm 0.33}$	$47.60_{\pm 2.25}$	$\mathbf{57.51}_{\pm 0.71}$	$\mathbf{77.03}_{\pm 0.49}$	$\mathbf{67.27}_{\pm 0.52}$

vidual components of DRIVE contribute effectively to the fine-tuning objectives? ❸ How sensitive is DRIVE to choices of essential hyper-parameters, including λ, τ, and η? ❹ Does diversity regularization accelerate convergence and enhance estimation robustness? ❺ Do loss-based contribution scores effectively guide weighting?

5.1 Setup

Baselines. We compare DRIVE with methods in the data-oblivious setting: Standard FT, linear probing, ℓ_2-regularization [38], WiSE-FT [79], LoRA [10], DEFT-UCS [22], FLOW [20], and TALR [6].

Datasets. Vision uses ImageNet-1K (anti-forgetting) and nine benchmarks: CIFAR-10/100 [80], Flowers102 [81], Caltech101 [82], Cars/Dogs [83, 84], Food101 [85], Beans [86], and EuroSAT [87]. LLMs fine-tune on MetaMathQA [88]/Magicoder [89] targeting GSM8K [90]/HumanEval [91], evaluated on MMLU [92], Commonsense, and MBPP [93]/GSM8K.

Evaluation metrics. For vision, we report fine-tuning accuracy and ImageNet-1K top-1 accuracy (restoring pre-trained heads) for anti-forgetting, alongside their average. For language, we employ `lm-evaluation-harness` [94] to assess target and general tasks, reporting the Overall metric (the average of target accuracy and general capability). Details in Appendix D.3.

Implementation details. For vision tasks, we use ImageNet-1K-pretrained ResNet-18 and ResNet-50 [95]. For language, we use Gemma-2-2B [96], Llama-3.2-3B and Llama-3.1-8B [97]. Across domains, we decay λ from 1 to 0, set τ to the 10th percentile of pre-training loss, $\eta = 1$, $k = 128$, $T_{\max} = 10$, and $\epsilon = 1e-5$; see Appendix D. Experiments used four A100 80GB GPUs.

5.2 Comparison on Vision and Language Tasks (Q1)

In the vision domain, where Standard FT often incurs severe catastrophic forgetting, DRIVE better reconciles adaptation and retention. As detailed in Table 2, Standard FT excels in adaptation but suffers severe forgetting; while FLOW attains competitive results overall average. Specifically, us-

ing ResNet-50, DRIVE achieves average accuracy gains of 2.0%, 1.54%, and 3.16% over FLOW on CIFAR-10, CIFAR-100, and Flowers102, respectively. These improvements are significant, stemming from higher target accuracy while largely retaining ImageNet-1K performance across benchmarks. Extensive experiments on six additional fine-grained datasets (Appendix E.1) corroborate these findings, showing consistent performance gains across diverse visual domains.

Table 3 shows that Standard FT achieves strong target performance but suffers from general capability degradation, dropping by up to 8.59% on Llama-3.1-8B. Retention-oriented baselines such as FLOW and TALR better preserve general capabilities, but lag behind in target accuracy. In contrast, DRIVE improves the general capability average over Standard FT by 5.12%, 4.87%, and 8.00% on Gemma-2-2B, Llama-3.2-3B, and Llama-3.1-8B, respectively, while maintaining competitive target accuracy. It ranks second on Gemma-2-2B and Llama-3.2-3B, and slightly surpasses Standard FT by 0.12% on Llama-3.1-8B.

These stability and plasticity gains yield the best overall performance across all three LLMs. DRIVE surpasses the strongest baseline by 1.68%, 1.43%, and 1.57% overall on Gemma-2-2B, Llama-3.2-3B, and Llama-3.1-8B, respectively. Overall, DRIVE secures high target performance while substantially reducing general capability degradation. Appendix E.3 extends our evaluation to a code dataset, and Appendix A.3 reports a replacement study on contribution scoring.

Table 4: DRIVE acts as a plug-and-play module, boosting performance in ℓ_2-Reg. and LoRA.

Method	Gen Acc.	Target Acc.	Overall Avg.
ℓ_2-Reg.	$55.58_{(+0.00)}$	$73.21_{(+0.00)}$	$64.40_{(+0.00)}$
+ FLOW	$56.08_{(+0.50)}$	$74.80_{(+1.59)}$	$65.44_{(+1.04)}$
+ **DRIVE**	$\mathbf{56.92}_{(+1.34)}$	$\mathbf{76.19}_{(+2.98)}$	$\mathbf{66.56}_{(+2.16)}$
LoRA	$56.90_{(+0.00)}$	$74.50_{(+0.00)}$	$65.70_{(+0.00)}$
+ FLOW	$57.52_{(+0.62)}$	$74.91_{(+0.41)}$	$66.22_{(+0.52)}$
+ **DRIVE**	$\mathbf{57.54}_{(+0.64)}$	$\mathbf{76.41}_{(+1.91)}$	$\mathbf{66.98}_{(+1.28)}$

DRIVE serves as a plug-and-play module that can be directly integrated with parameter-space fine-tuning algorithms. Table 4 validates DRIVE as a versatile module for parameter-space algorithms like ℓ_2-Regularization and LoRA on Llama-3.1-8B. In the ℓ_2-Regularization setting, DRIVE unlocks adaptation potential suppressed by rigid penalties, boosting target accuracy by 3.0% over the baseline, whereas FLOW improves it by 1.6%. For LoRA, DRIVE provides complementary regularization, surpassing LoRA+FLOW by 0.76% in overall average.

These stability and plasticity gains yield the best overall performance across all three LLMs. DRIVE exceeds the strongest baseline by 1.68, 1.43, and 1.57 percentage points on Gemma-2-2B, Llama-3.2-3B, and Llama-3.1-8B, respectively. To assess the statistical reliability of these improvements, we perform two-sided paired t-tests on the Overall scores from five runs with matched random seeds and data orders, using the existing training logs without additional training. All three improvements remain statistically significant after Holm correction across the three model-wise comparisons (adjusted $p \leq 0.0152$), and each individual 95% confidence interval for the mean paired difference lies entirely above zero. These results support the reliability of the observed gains under the evaluated matched-seed protocol. Appendix E.10 provides the complete analysis, including confidence intervals and paired effect sizes.

Overall, DRIVE maintains competitive target performance while substantially reducing general capability degradation. Appendix E.3 extends our evaluation to code fine-tuning, and Appendix A.3 reports a replacement study on contribution scoring.

5.3 Ablation Study (Q2)

Table 5 details the ablation results on Llama-3.1-8B using MetaMathQA. Removing any component degrades performance. Exclusively focusing on anti-forgetting or task adaptation fails to reach an optimal equilibrium; the former reduces target accuracy by 3.84%, while the latter degrades general capabilities by 5.27%. Likewise, removing uniform prior regularization causes the most severe impact, reducing average accuracy by 4.98%; see Appendix B.3 for its no-prior optimization. Removing diversity reg-

Table 5: Ablation study of DRIVE components.

Method	General	Target	Avg
DRIVE	57.51	77.03	**67.27**
- w/o Task-adaptation ($\lambda = 1$)	**58.42**	73.19	65.81
- w/o Anti-forgetting ($\lambda = 0$)	52.24	**79.83**	66.04
- w/o Diversity ($\eta = 0$)	56.18	74.78	65.48
- w/o KL Prior	53.96	70.61	62.29

ularization results in a 1.79% accuracy drop. Consistent trends also appear in vision tasks (Appendix E.8).

We examine whether DRIVE can benefit from an informative prior. Specifically, on Llama-3.1-8B using MetaMathQA, we compare the default uniform prior with an entropy-informed prior computed from the pre-trained model's predictive entropy. Let H_i denote the predictive entropy of sample $\boldsymbol{x}_i$. We assign larger prior mass to lower-entropy samples by setting $p_i \propto \exp(-H_i)$, followed by normalization over the training set. Table 6 shows that the entropy-informed prior achieves slightly better general and overall performance than the uniform prior, while maintaining comparable target accuracy. This suggests that DRIVE can benefit from simple auxiliary signals when available, while the uniform prior remains a strong default without additional estimators.

Table 6: Entropy vs. uniform prior.

Prior	General	Target	Avg.
Uniform Prior	57.51	**77.03**	67.27
Entropy Prior	**58.07**	76.94	**67.51**

Table 7: Diversity comparison.

Method	General	Target	Avg.
DEFT-UCS	54.11	74.37	64.24
DRIVE-Cluster	55.20	74.80	65.00
DRIVE	**57.51**	**77.03**	**67.27**

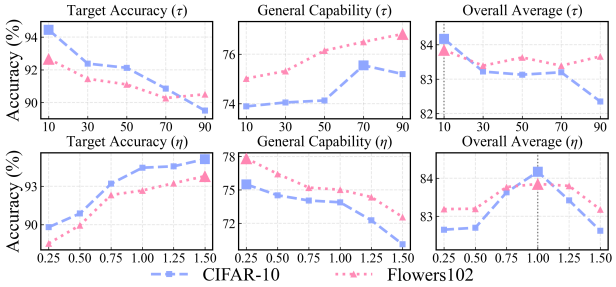


Figure 3: Sensitivity analysis of τ and η on ResNet-50.

To validate the effectiveness of our differentiable covariance diversity regularizer, we compare DRIVE against DEFT-UCS and a discrete variant, DRIVE-Cluster, on Llama-3.1-8B. We construct DRIVE-Cluster by replacing the differentiable diversity regularization with a discrete clustering mechanism. Specifically, it partitions data into $K = 7$ semantic groups. For a sample $\boldsymbol{x}_i$ assigned to the k-th cluster C_k, we adopt a balanced prior $p_i \equiv 1/(|C_k| \cdot K)$, to enforce equal aggregate cluster weights while determining within-cluster weights (details in Appendix F.2.2). As shown in Table 7, DRIVE-Cluster outperforms DEFT-UCS, highlighting the benefit of explicit diversity control. Crucially, DRIVE further surpasses DRIVE-Cluster by 2.27%, indicating that differentiable diversity regularization provides a more adaptive mechanism than rigid cluster balancing.

5.4 Hyperparameter Sensitivity Analysis (Q3)

To analyze how DRIVE mediates the stability-plasticity trade-off, we evaluate λ under static and dynamic settings. For static settings (0.0 to 1.0 at intervals of 0.2), Table 8 on ResNet-50 shows that lower values favor plasticity but worsen forgetting, whereas higher values prioritize retention, achieving an optimal static balance at $\lambda = 0.8$. Among dynamic schedules, the decaying strategy ($1 \rightarrow 0$) yields superior average accuracy, outperforming the best static configuration ($\lambda = 0.8$).

Regarding hyperparameters in Fig. 3, τ controls weight sharpness through smoothing. Small values favor target adaptation, moderate values give the best overall trade-off, and overly large values make weights nearly uniform and can weaken retention. Increasing η improves target performance up to $\eta = 1.0$, after which generalization degrades.

5.5 Analysis of Convergence and Robustness (Q4)

We investigate the optimization dynamics induced by diversity regularization and evaluate estimation robustness by analyzing convergence curves, measuring pairwise parameter distances across four seeds, and examining performance on the top 10% samples with the highest pre-training loss.

Fig. 4 illustrates diversity term's impact on training dynamics using Llama-3.1-8B on MetaMathQA. We define convergence as the point where the relative fluctuation of the smoothed loss remains within 2% over a 400-step window. DRIVE stabilizes significantly earlier than baselines lacking diversity constraints. This supports our claim that diversity accelerates convergence.

To verify DRIVE mitigates the under-fitting inherent in stability-oriented methods like FLOW, we further tested ResNet-50 on the most challenging 10% samples by pre-training loss. In Fig. 5, FLOW performs poorly on these challenging samples. Standard FT excels but suffers catastrophic forgetting

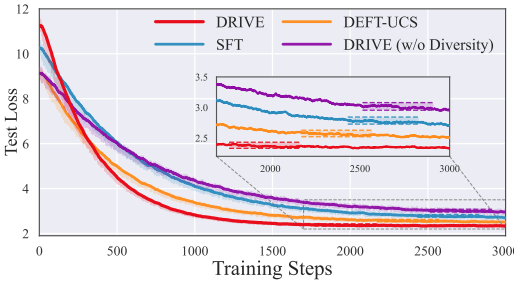


Figure 4: Convergence curves across baselines.

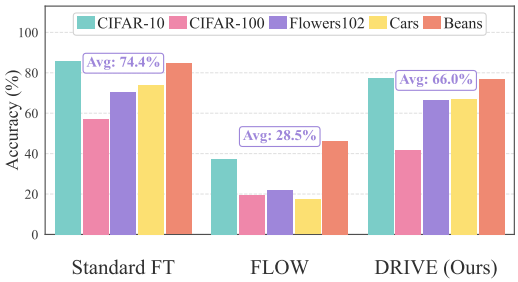


Figure 5: Hard-sample accuracy across datasets

Table 8: Sensitivity Analysis of λ.

	CIFAR-10			Flowers102		
λ	IN-1K	Target	Avg.	IN-1K	Target	Avg.
0.0	61.84	**97.10**	79.47	71.98	**93.67**	82.83
0.2	64.92	96.08	80.50	72.47	92.63	82.55
0.4	67.31	95.21	81.26	73.39	92.55	82.97
0.6	70.28	94.63	82.46	75.06	91.53	83.30
0.8	73.02	94.19	83.61	75.42	91.69	83.56
1.0	**76.10**	90.02	83.06	**76.26**	90.73	83.50
$0 \to 1$	70.95	94.88	82.92	73.67	93.01	83.34
$1 \to 0$	73.89	94.45	**84.17**	75.02	92.67	**83.85**

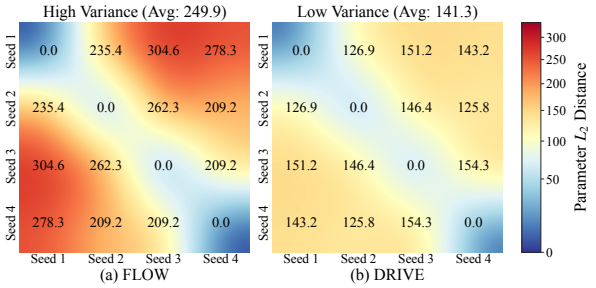


Figure 6: Pairwise parameter distance heatmap.

on the retention set. Conversely, DRIVE maintains satisfactory performance on difficult samples due to its explicit contribution weighting mechanism. More detailed analysis is provided in Appendix F.4.

To quantify estimation robustness, we analyzed geometric consistency across four random seeds via the Euclidean (L_2) distance between flattened fine-tuned parameter vectors. As shown in Fig. 6 for ResNet-50 on CIFAR-100, FLOW exhibits volatility and pronounced seed sensitivity. In contrast, DRIVE enhances geometric stability, reducing the average pairwise parameter distance by 43.46%. This contraction further confirms that our diversity regularization imposes an effective constraint, minimizing the estimator's generalized variance and yielding more reliable parameter estimates.

5.6 Validation of Loss-Based Contribution Scores (Q5)

On ResNet-50/Flowers102 (1,020 training samples; $\lambda = 0.8$), c_i correlates with validation-gradient-based Shapley approximations [71], achieving Spearman's $\rho = 0.83$ and Pearson's $r = 0.81$ (both $p < 0.001$). Reference validation sets are used only for diagnostics (Appendix A.2). In a single-seed diagnostic on Llama-3.1-8B/MetaMathQA, shuffling or reversing score rankings while preserving score values and other settings reduces Overall by 2.17 and 7.06 percentage points, respectively (Table 9), supporting the utility of sample-specific contribution scores.

Table 9: Effect of score perturbations.

Setting	General	Target	Overall
DRIVE	**57.48**	**77.10**	**67.29**
Shuffled	56.31	73.92	65.12
Reversed	50.02	70.43	60.23

6 Conclusion

We presented DRIVE, a novel data weighting framework for data-oblivious fine-tuning. Rather than aggregating off-the-shelf signals, DRIVE recasts contribution, diversity, and sample prior into a single coherent differentiable objective. Concretely, contribution is partitioned into retention and adaptation via a tunable Shapley-additive decomposition, diversity is captured through a differentiable covariance-determinant term, and weights are anchored by a prior regularizer. In linear and local nonlinear regimes, this unified modeling tightens the stability–plasticity trade-off. Diversity provably accelerates convergence and bolsters estimation robustness; the solver converges linearly. Across vision and language benchmarks, DRIVE consistently surpasses SOTA methods. Future work explores samples beneficial to both contributions and token-level extensions.

Acknowledgments and Disclosures

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62476191 and 42550187, with computing resources provided by our research group. The authors declare no competing interests.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR 2016*, 2016, pp. 770–778.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*, 2019, pp. 4171–4186.

[3] C.-C. Wu, Z. R. Tam, C.-Y. Lin, Y.-N. Chen, S.-H. Sun, and H. yi Lee, "Mitigating forgetting in LLM fine-tuning via low-perplexity token learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=1ktdvp1EYI

[4] Y. Wang, A. Bai, N. Peng, and C.-J. Hsieh, "On the loss of context awareness in general instruction fine-tuning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. [Online]. Available: https://openreview.net/forum?id=7OdU0LYXLr

[5] J. Harmon, A. Hochlehnert, M. Bethge, and A. Prabhu, "Mapping post-training forgetting in language models at scale," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=qCIg2WGudx

[6] J. Lin, Z. Wang, K. Qian, T. Wang, A. Srinivasan, H. Zeng, R. Jiao, X. Zhou, J. Gesi, D. Wang, Y. Guo, K. Zhong, W. Zhang, S. Sanghavi, C. Chen, H. Yun, and L. Li, "SFT doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in LLMs," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=ru93xpQFi1

[7] E. S. Lubana, P. Trivedi, D. Koutra, and R. Dick, "How do quadratic regularizers prevent catastrophic forgetting: The role of interpolation," in *Conference on Lifelong Learning Agents*. PMLR, 2022, pp. 819–837.

[8] Y. Chen, S. Wang, Y. Zhang, Z. Lin, H. Zhang, W. Sun, T. Ding, and R. Sun, "Mofo: Momentum-filtered optimizer for mitigating forgetting in llm fine-tuning," *arXiv preprint arXiv:2407.20999*, 2024.

[9] A. Panda, B. Isik, X. Qi, S. Koyejo, T. Weissman, and P. Mittal, "Lottery ticket adaptation: Mitigating destructive interference in LLMs," in *ICML 2024 Next Generation of AI Safety Workshop*, 2024. [Online]. Available: https://openreview.net/forum?id=AfdAZtUWsR

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[11] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, "LoRA learns less and forgets less," *Transactions on Machine Learning Research*, 2024, featured Certification. [Online]. Available: https://openreview.net/forum?id=aloEru2qCG

[12] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. Farhadi, "Editing models with task arithmetic," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=6t0Kwf8-jrj

[13] A. Kleiman, G. K. Dziugaite, J. Frankle, S. Kakade, and M. Paul, "Soup to go: mitigating forgetting during continual learning with model averaging," *arXiv preprint arXiv:2501.05559*, 2025.

[14] W. Li, L. Zou, M. Tang, Q. Yu, W. Li, and C. Li, "Meta-lora: Memory-efficient sample reweighting for fine-tuning large language models," in *Proceedings of the 31st international conference on computational linguistics*, 2025, pp. 8504–8517.

[15] M. Thakkar, T. Bolukbasi, S. Ganapathy, S. Vashishth, S. Chandar, and P. Talukdar, "Self-influence guided data reweighting for language model pre-training," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2033–2045. [Online]. Available: https://aclanthology.org/2023.emnlp-main.125/

[16] D. Sow, H. Woisetschläger, S. Bulusu, S. Wang, H. A. Jacobsen, and Y. Liang, "Dynamic loss-based sample reweighting for improved large language model pretraining," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=gU4ZgQNsOC

[17] W. Zhao, Y. Chen, Y. Tang, W. Ma, S. Hu, S. X. Hu, A. Iacob, A. Mehrotra, and N. D. Lane, "Rethinking data curation in LLM training: Online reweighting offers better generalization than offline methods," in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=UFwnsmFZ6R

[18] X. Zhou, O. Wu, W. Zhu, and Z. Liang, "Understanding difficulty-based sample weighting with a universal difficulty measure," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 68–84.

[19] J. Shu, X. Yuan, D. Meng, and Z. Xu, "Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11 521–11 539, 2023.

[20] S. Sanyal, H. Prairie, R. Das, A. Kavis, and S. Sanghavi, "Upweighting easy samples in fine-tuning mitigates forgetting," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=13HPTmZKbM

[21] W. Liu, W. Zeng, K. He, Y. Jiang, and J. He, "What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=BTKAeLqLMw

[22] D. Das and V. Khetan, "Deft-ucs: Data efficient fine-tuning for pre-trained language models via unsupervised core-set selection for text-editing," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 20 296–20 312.

[23] W. Li, C. Chen, and P. Varakantham, "Unlocking the planning capabilities of large language models with maximum diversity fine-tuning," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 3318–3340.

[24] B. Zhang, J. Wang, Q. Du, J. Zhang, Z. Tu, and D. Chu, "A survey on data selection for llm instruction tuning," *Journal of Artificial Intelligence Research*, vol. 83, 2025.

[25] P. Wang, Y. Shen, Z. Guo, M. Stallone, Y. Kim, P. Golland, and R. Panda, "Diversity measurement and subset selection for instruction tuning datasets," in *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2025. [Online]. Available: https://openreview.net/forum?id=cV9OF45hBb

[26] L. Liu, X. Liu, D. F. Wong, D. Li, Z. Wang, B. Hu, and M. Zhang, "Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection," *Advances in Neural Information Processing Systems*, vol. 37, pp. 97 800–97 825, 2024.

[27] Y. Li, B. Hui, X. Xia, J. Yang, M. Yang, L. Zhang, S. Si, L.-H. Chen, J. Liu, T. Liu, F. Huang, and Y. Li, "One-shot learning as instruction data prospector for large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 4586–4601. [Online]. Available: https://aclanthology.org/2024.acl-long.252/

[28] M. Li, Y. Zhang, Z. Li, J. Chen, L. Chen, N. Cheng, J. Wang, T. Zhou, and J. Xiao, "From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 7602–7635. [Online]. Available: https://aclanthology.org/2024.naacl-long.421/

[29] J. Luo, X. Luo, X. Chen, Z. Xiao, W. Ju, and M. Zhang, "Semi-supervised fine-tuning for large language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 2795–2808.

[30] J. Zhang, C.-X. Zhang, Y. Liu, Y.-X. Jin, X.-W. Yang, B. Zheng, Y. Liu, and L.-Z. Guo, "D3: diversity, difficulty, and dependability-aware data selection for sample-efficient llm instruction tuning," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 8348–8356.

[31] Z. Han, Z. Liang, F. Yang, L. Liu, L. Li, Y. Bian, P. Zhao, B. Wu, C. Zhang, and J. Yao, "Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup," *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 704–37 718, 2022.

[32] R. Zhang, C. Zhang, X. Zhang, L. Qiu, H. Jiang, Y. Zhuang, Q. Zhang, H. Yun, X. Li, B. Yin *et al.*, "Dorm: Preference data weights optimization for reward modeling in llm alignment," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 22 721–22 739.

[33] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. Lee, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 5254–5276.

[34] X.-K. Wu, M. Chen, W. Li, R. Wang, L. Lu, J. Liu, K. Hwang, Y. Hao, Y. Pan, Q. Meng *et al.*, "Llm fine-tuning: Concepts, opportunities, and challenges," *Big Data and Cognitive Computing*, vol. 9, no. 4, p. 87, 2025.

[35] W. Kuang, B. Qian, Z. Li, D. Chen, D. Gao, X. Pan, Y. Xie, Y. Li, B. Ding, and J. Zhou, "Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5260–5271.

[36] X. Chen, L. Yang, and O. Wu, "One algorithm, two goals: Dual scoring for parameter and data selection in llm fine-tuning," 2026. [Online]. Available: https://arxiv.org/abs/2605.06166

[37] J. Zhang, O. Wu, and Y. Du, "Nads: Navigator-guided data selection for mitigating catastrophic forgetting in fine-tuning," 2026.

[38] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[39] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International conference on machine learning*. PMLR, 2018, pp. 4528–4537.

[40] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 139–154.

[41] J. Lee, H. G. Hong, D. Joo, and J. Kim, "Continual learning with extended kronecker-factored approximate curvature," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9001–9010.

[42] J. Sliwa, F. Schneider, P. Hennig, and J. M. Hernández-Lobato, "Mitigating forgetting in low rank adaptation," in *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025. [Online]. Available: https://openreview.net/forum?id=W3caXQAjcu

[43] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-GEM," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Hkf2_sC5FX

[44] G. Zeng, Y. Chen, B. Cui, and S. Yu, "Continual learning of context-dependent processing in neural networks," *Nature Machine Intelligence*, vol. 1, no. 8, pp. 364–372, 2019.

[45] M. Farajtabar, N. Azizan, A. Mott, and A. Li, "Orthogonal gradient descent for continual learning," in *International conference on artificial intelligence and statistics*. PMLR, 2020, pp. 3762–3773.

[46] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.

[47] L. Wang, X. Zhang, Q. Li, M. Zhang, H. Su, J. Zhu, and Y. Zhong, "Incorporating neuro-inspired adaptability for continual learning in artificial intelligence," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1356–1368, 2023.

[48] M. B. Gurbuz and C. Dovrolis, "Nispa: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8157–8174.

[49] S. Jung, H. Ahn, S. Cha, and T. Moon, "Continual learning with node-importance based adaptive group sparse regularization," *Advances in neural information processing systems*, vol. 33, pp. 3647–3658, 2020.

[50] R. Aljundi, M. Lin, B. Goujaud, and Y. Bengio, "Gradient based sample selection for online continual learning," *Advances in neural information processing systems*, vol. 32, 2019.

[51] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi, "Rainbow memory: Continual learning with a memory of diverse samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8218–8227.

[52] R. Tiwari, K. Killamsetty, R. Iyer, and P. Shenoy, "Gcr: Gradient coreset based replay buffer selection for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 99–108.

[53] M. De Lange and T. Tuytelaars, "Continual prototype evolution: Learning online from non-stationary data streams," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8250–8259.

[54] X. Chen, Z. Wang, D. Sow, J. Yang, T. Chen, Y. Liang, M. Zhou, and Z. Wang, "Take the bull by the horns: Hard sample-reweighted continual training improves llm generalization," *arXiv preprint arXiv:2402.14270*, 2024.

[55] S. Mukherjee, V. D. Lai, R. Addanki, R. Rossi, S. Yoon, T. Bui, A. Rao, J. Subramanian, and B. Kveton, "Learning to clarify by reinforcement learning through reward-weighted fine-tuning," *arXiv preprint arXiv:2506.06964*, 2025.

[56] H.-Y. Kuo, Y.-H. Liao, Y.-C. Chao, W.-Y. Ma, and P.-J. Cheng, "Not all LLM-generated data are equal: Rethinking data weighting in text classification," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=oI5tZaWkF9

[57] J. Han, H. Liu, J. Fang, N. Tan, and H. Xiong, "Automatic instruction data selection for large language models via uncertainty-aware influence maximization," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 4969–4979.

[58] F. Kang, H. A. Just, Y. Sun, H. Jahagirdar, Y. Zhang, R. Du, A. K. Sahu, and R. Jia, "Get more for less: Principled data selection for warming up fine-tuning in LLMs," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=QmYNBVukex

[59] X. Zeng, H. Wang, J. Lin, J. Wu, T. Cody, and D. Zhou, "LensLLM: Unveiling fine-tuning dynamics for LLM selection," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=om0CcjvEQh

[60] Y. Ju, Z. Ni, X. Xing, Z. Zeng, H. Zhao, S. Fan, and Z. Zhang, "Mitigating training imbalance in llm fine-tuning via selective parameter merging," in *EMNLP*, 2024.

[61] F. P. C. Nuti, T. Franzmeyer, and J. F. Henriques, "Tuco: Measuring the contribution of fine-tuning to individual responses of LLMs," in *Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=yxoxw0IUTR

[62] D. Stap and C. Monz, "The effect of language diversity when fine-tuning large language models for translation," *arXiv preprint arXiv:2505.13090*, 2025.

[63] Q. Wang, S. Pan, T. Linzen, and E. Black, "Multilingual prompting for improving llm generation diversity," *arXiv preprint arXiv:2505.15229*, 2025.

[64] Z. Li, C. Chen, T. Xu, Z. Qin, J. Xiao, R. Sun, and Z.-Q. Luo, "Entropic distribution matching for supervised fine-tuning of llms: Less overfitting and better diversity," in *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024.

[65] D. Zhang, J. Wang, and F. Charton, "Diversification catalyzes language models' instruction generalization to unseen semantics," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 23 236–23 249.

[66] C. Chai, J. Liu, N. Tang, J. Fan, D. Miao, J. Wang, Y. Luo, and G. Li, "Goodcore: Data-effective and data-efficient machine learning through coreset selection over incomplete data," *Proceedings of the ACM on Management of Data*, vol. 1, no. 2, pp. 1–27, 2023.

[67] H. Li, X. Li, Y. Dong, and K. Liu, "From macro to micro: Probing dataset diversity in language model fine-tuning," *arXiv preprint arXiv:2505.24768*, 2025.

[68] F. Song, B. Yu, H. Lang, H. Yu, F. Huang, H. Wang, and Y. Li, "Scaling data diversity for fine-tuning language models in human alignment," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 14 358–14 369.

[69] Y. Chen, L. Wolf, S. Chakraborty, A. Pacchiano, and I. Paschalidis, "Enhancing diversity in large language models via determinantal point processes," in *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025. [Online]. Available: https://openreview.net/forum?id=v74LJpeUYX

[70] L. S. Shapley, "A value for n-person games," *Annals of Mathematics Studies*, vol. 28, pp. 307–317, 1953.

[71] J. T. Wang, P. Mittal, D. Song, and R. Jia, "Data shapley in one training run," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=HD6bWcj87Y

[72] K. Chaloner and I. Verdinelli, "Bayesian experimental design: A review," *Statistical science*, pp. 273–304, 1995.

[73] A. Alexanderian, "Optimal experimental design for infinite-dimensional bayesian inverse problems governed by pdes: A review," *Inverse Problems*, vol. 37, no. 4, p. 043001, 2021.

[74] S. Chen, Y. Hou, Y. Cui, W. Che, T. Liu, and X. Yu, "Recall and learn: Fine-tuning deep pretrained language models with less forgetting," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7870–7881. [Online]. Available: https://aclanthology.org/2020.emnlp-main.634/

[75] S. Wei, C. Luo, and Y. Luo, "Scaled decoupled distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 975–15 983.

[76] S. Khosla, A. Tiwari, K. Kafle, S. Jenni, H. Zhao, J. Collomosse, and J. Shi, "Magnet: Augmenting generative decoders with representation learning and infilling capabilities," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 27 328–27 346.

[77] A. Deshpande, A. Achille, A. Ravichandran, H. Li, L. Zancato, C. Fowlkes, R. Bhotika, S. Soatto, and P. Perona, "A linearized framework and a new benchmark for model selection for fine-tuning," *arXiv preprint arXiv:2102.00084*, 2021.

[78] A. Tomihari and I. Sato, "Understanding linear probing then fine-tuning language models from ntk perspective," *Advances in Neural Information Processing Systems*, vol. 37, pp. 139 786–139 822, 2024.

[79] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong *et al.*, "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7959–7971.

[80] A. Krizhevsky and G. E. Hinton, "Learning multiple layers of features from tiny images," in *Technical report, University of Toronto*, 2009.

[81] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.

[82] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[83] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[84] A. Khosla, N. Jayadevaprakash, B. Yao, and F.-F. Li, "Novel dataset for fine-grained image categorization: Stanford dogs," in *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2, no. 1, 2011.

[85] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European conference on computer vision*. Springer, 2014, pp. 446–461.

[86] E. Elfatimi, R. Eryigit, and L. Elfatimi, "Beans leaf diseases classification using mobilenet models," *Ieee Access*, vol. 10, pp. 9471–9482, 2022.

[87] P. Helber, B. Bischke, A. Dengel, and D. 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*, vol. 12, no. 7, pp. 2217–2226, 2019.

[88] L. Yu, W. Jiang, H. Shi, J. YU, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu, "Metamath: Bootstrap your own mathematical questions for large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=N8N0hgNDRt

[89] Y. Wei, Z. Wang, J. Liu, Y. Ding, and L. ZHANG, "Magicoder: Empowering code generation with OSS-instruct," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=XUeoOBid3x

[90] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[91] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[92] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=d7KBjmI3GmQ

[93] J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le *et al.*, "Program synthesis with large language models," *arXiv preprint arXiv:2108.07732*, 2021.

[94] L. Gao, J. Tow, B. Abbasi, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, A. Le Noac'h *et al.*, "A framework for few-shot language model evaluation," *URL https://zenodo. org/records/12608602*, 2024.

[95] R. Wightman, H. Touvron, and H. Jegou, "Resnet strikes back: An improved training procedure in timm," in *NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future*, 2021. [Online]. Available: https://openreview.net/forum?id=NG6MJnVl6M5

[96] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé *et al.*, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[97] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[98] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[99] M. Paul, S. Ganguli, and G. K. Dziugaite, "Deep learning on a data diet: Finding important examples early in training," *Advances in neural information processing systems*, vol. 34, pp. 20 596–20 607, 2021.

[100] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=VD-AYtP0dve

[101] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[102] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[103] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "Hellaswag: Can a machine really finish your sentence?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4791–4800.

[104] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.

[105] Y. Bisk, R. Zellers, J. Gao, Y. Choi *et al.*, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 7432–7439.

[106] T. Mihaylov, P. Clark, T. Khot, and A. Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2381–2391.

[107] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[108] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[109] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen, "LESS: Selecting influential data for targeted instruction tuning," in *Forty-first International Conference on Machine Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=PG5fV50maR

[110] H. Ritter, A. Botev, and D. Barber, "A scalable laplace approximation for neural networks," in *6th international conference on learning representations, ICLR 2018-conference track proceedings*, vol. 6. International Conference on Representation Learning, 2018.

Organization of the Appendix

A More Details Supporting Our Contribution Function C 18
A.1 Inapplicable Contribution Measures 18
A.2 Statistical Support for the Loss-Based Contribution Proxies 18
A.3 Alternative Validation-Free Metrics 19
B Solution for Objective (4) and Convergence Analysis 20
B.1 Concavity and Uniqueness of Eq. (4) 20
B.2 Derivation of the Iterative Update Rule 21
B.3 Optimization for the No-Prior Ablation 22
B.4 Convergence of the Weight Solver 22
B.5 Justification for Randomized Dimensionality Reduction 23
C Detailed Theoretical Proofs 23
C.1 Proof for the Linear Regime in Sec. 4.1 23
C.2 Local Nonlinear Analysis with Hessian Remainder Control 25
D More Experimental Details 26
D.1 Baseline Details 26
D.2 Vision Datasets and Evaluation Details 27
D.3 LLM Datasets and Evaluation Details 27
D.4 Implementation Details 28
E Additional Experimental Results and Analysis 28
E.1 Analysis of Vision Tasks 29
E.2 Experiments on Vision Transformers 29
E.3 Experiments on Language Models 30
E.4 Impact of Balance Parameter λ on Language Tasks 30
E.5 Temporal Dynamics of λ: Increasing vs. Decaying Schedule 31
E.6 Impact of Prior Regularization τ 31
E.7 Impact of Diversity Regularization η on Language Tasks 32
E.8 Ablation Study on Vision Tasks 32
E.9 Computational Efficiency Analysis 33
F Analysis of Contribution Assessment and Diversity Mechanisms 34
F.1 Shapley Analysis of Anti-Forgetting vs. Task Adaptation 34
F.2 Sample Diversity 35
F.3 Analysis of Covariance Approximation Strategies 37
F.4 Empirical Verification of Dynamic Adaptation on Hard Samples 38
F.5 Sensitivity to the Projection Dimension k 38
G Limitations 38

A More Details Supporting Our Contribution Function C

This appendix provides additional support for the contribution function C. Sec. A.1 explains why exact Shapley values and validation-dependent contribution estimates cannot be directly deployed under our strict data-oblivious setting. Sec. A.2 shows that the adopted loss-based proxies remain statistically aligned with validation-based Shapley-style reference contributions in the aspects most relevant to soft weighting. Sec. A.3 discusses several practical validation-free alternatives and empirically compares representative variants, showing that the standard loss-based instantiation offers the best effectiveness–efficiency trade-off.

A.1 Inapplicable Contribution Measures

A natural theoretical foundation for sample contribution is the Shapley value [70]. Let $U(\cdot)$ denote a utility function and D be a training set of N samples. For a sample $z \in D$, its Shapley value is

$$\phi_z(U) := \frac{1}{N} \sum_{k=1}^{N} \binom{N-1}{k-1}^{-1} \sum_{\substack{S \subseteq D \setminus \{z\} \\ |S|=k-1}} [U(S \cup \{z\}) - U(S)] \,. \tag{6}$$

This provides a rigorous conceptual foundation for contribution modeling. However, exact Shapley computation requires repeated retraining over exponentially many subsets, which is infeasible for large-scale fine-tuning. Thus, exact Shapley values can only serve as a theoretical reference, but not as the deployed contribution measure in DRIVE.

To avoid exhaustive retraining, recent work has proposed one-run approximations that estimate contribution by accumulating gradient alignment along a single training trajectory [71]. In this case, one define validation-based reference contributions using retain and target validation gradients:

$$a_i^{\text{ref}} := \sum_{t=1}^{T} \eta_t \left\langle \nabla_{\theta_t} \ell(f(x_i;\theta_t), y_i), \nabla_{\theta_t} \ell(f(D_{\text{val}}^{\text{retain}};\theta_t), y_{\text{val}}^{\text{retain}}) \right\rangle, \tag{7}$$

$$t_i^{\text{ref}} := \sum_{t=1}^{T} \eta_t \left\langle \nabla_{\theta_t} \ell(f(x_i;\theta_t), y_i), \nabla_{\theta_t} \ell(f(D_{\text{val}}^{\text{target}};\theta_t), y_{\text{val}}^{\text{target}}) \right\rangle. \tag{8}$$

These quantities are much more tractable than exact Shapley values and are useful as diagnostic references. However, they still require two external supervision channels: a retain validation set and a target validation set. Similarly, meta-learning-based contribution learning [98] also depends on explicit validation supervision. In contrast, DRIVE is designed for a strict data-oblivious fine-tuning setting, where neither the original pre-training data nor auxiliary retain/target validation sets are assumed to be available during training.

Under the strict data-oblivious setting considered by DRIVE, more principled contribution measures are either computationally intractable (exact Shapley) or fundamentally incompatible with the available supervision (validation-based approximations and meta-learning). This motivates our use of a training-time computable, validation-free proxy in the main text.

A.2 Statistical Support for the Loss-Based Contribution Proxies

Although the loss-based definitions in Eq. (3) are adopted for practicality, the question is whether they preserve the information that matters for weighting. Since Eq. (5) depends on the relative ordering and separation of contribution scores, a proxy should preserve trend, ranking structure, head-sample identity, and the agreement of the unified contribution score with its reference counterpart. In this experiment, the unified score is computed with the same setting as the main text, $c_i = \lambda a_i + (1-\lambda) t_i$ with $\lambda = 0.8$, and $c_i^{\text{ref}} = \lambda a_i^{\text{ref}} + (1-\lambda) t_i^{\text{ref}}$. Following the validation-based Shapley-style formulation in [71], we use the sample contributions computed on the retain validation set and the target validation set as reference anti-forgetting and task-adaptation contributions, denoted by a_i^{ref} and t_i^{ref}, respectively. These reference quantities are used only for evaluating the quality of the loss-based proxies.

We therefore report four complementary statistics. Quantile Monotonicity Score measures global trend preservation by sorting samples with the proxy score, partitioning them into equal-frequency

Table 10: Agreement between loss-based proxy contributions and validation-based Shapley reference scores on ResNet-50 fine-tuned on Flowers102, using an ImageNet subset as the retain validation set.

Metric	**Anti-forgetting**	**Task adaptation**	**Unified score c_i**
Quantile Monotonicity Score ↑	0.89	0.98	0.92
Spearman ρ ↑	0.76	0.87	0.83
Top-50% Overlap ↑	0.87	0.95	0.89
Pearson r ↑	0.65	0.87	0.81

quantile bins, and computing the fraction of adjacent bins whose mean reference score is monotonically non-decreasing. Top-50% Overlap measures head-sample consistency as the overlap between the top half selected by the proxy and by the reference. Spearman's ρ evaluates ranking agreement, while Pearson's r serves as a supplementary linear amplitude-based statistic. All reported results are highly significant ($p < 0.001$ for every metric under the corresponding tests).

Across these decision-oriented evaluations, the loss-based instantiation used in DRIVE shows strong agreement with the validation-based Shapley-style reference contributions. The quantile trends remain clearly monotone, the Top-50% overlap is high for the influential samples that dominate weighting, and both Spearman and Pearson correlations are strong. Importantly, this conclusion holds not only for the decomposed anti-forgetting and task-adaptation components, but also for the unified contribution score c_i, which is the quantity directly used by DRIVE to determine sample weights.

A.3 Alternative Validation-Free Metrics

While the main text adopts loss-based contributions, the DRIVE framework can accommodate other practical validation-free surrogates. Below we discuss several representative alternatives. Here, L denotes the number of prediction positions over which the surrogate score is averaged; for LLM fine-tuning, it is the sequence length, while for vision classification we set $L = 1$.

Gradient-based metrics. The gradient norm provides a more direct first-order measure of sample-level update influence than scalar loss [99]. A natural alternative is

$$a_i = -\left\|\nabla_{\theta_{\text{pre}}} \ell\big(f(x_i;\theta_{\text{pre}}), y_i\big)\right\|_2, \tag{9}$$

$$t_i = \left\|\nabla_{\theta} \ell\big(f(x_i;\theta), y_i\big)\right\|_2. \tag{10}$$

This formulation is theoretically appealing and closely connected to GraNd-style sample valuation [99], but it requires per-sample backward computation and is significantly more expensive in large-scale fine-tuning.

Error-based metrics. EL2N measures the ℓ_2 discrepancy between the predicted probability vector and the target one-hot vector [99]:

$$\text{EL2N}(x_i;\theta) = \frac{1}{L}\sum_{j=1}^{L}\left\|p_j(x_i;\theta) - y_j\right\|_2, \tag{11}$$

based on which we define

$$a_i = -\text{EL2N}(x_i;\theta_{\text{pre}}), \qquad t_i = \text{EL2N}(x_i;\theta). \tag{12}$$

This yields a bounded and relatively robust signal, but in practice it is less directly aligned with optimization dynamics than the adopted loss-based formulation.

Uncertainty-based metrics. Prediction entropy provides another alternative, motivated by semantic uncertainty [100]:

$$a_i = -\frac{1}{L}\sum_{j=1}^{L}\left(-\sum_{k=1}^{|V|} p_{j,k}(x_i;\theta_{\text{pre}})\log p_{j,k}(x_i;\theta_{\text{pre}})\right), \tag{13}$$

$$t_i = \frac{1}{L}\sum_{j=1}^{L}\left(-\sum_{k=1}^{|V|} p_{j,k}(x_i;\theta)\log p_{j,k}(x_i;\theta)\right). \tag{14}$$

Entropy-based scoring emphasizes uncertainty, but it may over-focus on ambiguous or difficult samples and is not always as stable for balancing retention and adaptation.

Table 11: Replacement study of practical contribution mechanisms on Llama-3.2 3B. Gradient-based scoring is slightly stronger in accuracy but substantially more expensive, while the standard loss-based DRIVE offers the best effectiveness–efficiency trade-off.

Method	General Capability Acc. (%)				Target Acc. (%)	Overall	Relative Time
	CS	MMLU	MBPP	Avg	GSM8K	Avg (%)	Cost
DRIVE-Gra	55.11	**55.94**	**41.48**	50.84	**65.13**	**57.99**	2.12×
DRIVE-Entro	55.14	55.40	40.40	50.31	59.97	55.14	1.04×
DRIVE	**55.92**	55.71	41.00	**50.88**	64.50	57.69	1.00×

Margin-based metrics. A margin-based alternative is

$$a_i = \frac{1}{L}\sum_{j=1}^{L}\Big(z_{j,y_{i,j}}(x_i;\theta_{\text{pre}}) - \max_{k\neq y_{i,j}} z_{j,k}(x_i;\theta_{\text{pre}})\Big), \tag{15}$$

$$t_i = -\frac{1}{L}\sum_{j=1}^{L}\Big(z_{j,y_{i,j}}(x_i;\theta) - \max_{k\neq y_{i,j}} z_{j,k}(x_i;\theta)\Big), \tag{16}$$

which prioritizes boundary-near target samples while favoring confidently mastered prior knowledge.

Replacement study. To verify that the loss-based instantiation is a reasonable default, we compare DRIVE against two representative alternatives: a gradient-norm variant (DRIVE-Gra) and an entropy-based variant (DRIVE-Entro). As shown in Table 11, DRIVE-Gra achieves the highest overall score, indicating that richer gradient signals can improve contribution estimation. However, this gain comes at a computational cost, requiring roughly twice the training time of standard DRIVE. By contrast, DRIVE-Entro underperforms both. Notably, the loss-based DRIVE remains competitive in accuracy while offering the strongest efficiency profile. This supports our choice of the loss-based contribution function as the default practical instantiation of C.

B Solution for Objective (4) and Convergence Analysis

We analyze the optimization and efficient implementation of Eq. (4). Sec. B.1 proves concavity and uniqueness, Sec. B.2 derives the KL-regularized iterative update rule, Sec. B.3 explains how the no-prior ablation is optimized, Sec. B.4 establishes solver convergence, and Sec. B.5 justifies the randomized projection used for efficient covariance inversion.

B.1 Concavity and Uniqueness of Eq. (4)

Recall the unified weight optimization objective introduced in Eq. (4). Following the implementation and the solver analysis, we use the ridge-regularized weighted covariance $\mathbf{\Sigma}_{\boldsymbol{w}}$, where the ridge term is absorbed into the notation. The objective is

$$\mathcal{J}(\boldsymbol{w}) = \underbrace{\sum_{i=1}^{N} w_i c_i}_{\mathcal{J}_{\text{linear}}} \underbrace{- \tau \sum_{i=1}^{N} w_i \log \frac{w_i}{p_i}}_{\mathcal{J}_{\text{KL}}} + \underbrace{\eta \log |\mathbf{\Sigma}_{\boldsymbol{w}}|}_{\mathcal{J}_{\text{div}}}, \tag{17}$$

subject to the simplex constraint

$$\Delta^{N-1} = \left\{ \boldsymbol{w} \in \mathbb{R}^N \mid \sum_{i=1}^{N} w_i = 1,\ w_i \geq 0 \right\}.$$

We assume $p_i > 0$ for all i, $\tau > 0$, and $\eta \geq 0$.

Proposition 1 (Strict Concavity and Uniqueness)**.** *The objective $\mathcal{J}(\boldsymbol{w})$ is strictly concave on the simplex. Consequently, Eq. (4) admits a unique global maximizer $\boldsymbol{w}^*$.*

Proof. The linear term $\mathcal{J}_{\text{linear}}$ is affine in $\boldsymbol{w}$, so it is both convex and concave. The KL term satisfies

$$D_{\text{KL}}(\boldsymbol{w}\|\boldsymbol{p}) = \sum_{i=1}^{N} w_i \log w_i - \sum_{i=1}^{N} w_i \log p_i.$$

Since $x \log x$ is strictly convex on $x > 0$, $D_{\text{KL}}(\boldsymbol{w}\|\boldsymbol{p})$ is strictly convex on the relative interior of the simplex. Therefore, $-\tau D_{\text{KL}}(\boldsymbol{w}\|\boldsymbol{p})$ is strictly concave for $\tau > 0$. Equivalently, its Hessian on the positive orthant is diagonal with entries $-\tau/w_i$, which is negative definite.

The diversity term is also concave. Since the ridge-regularized covariance $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ is affine in $\boldsymbol{w}$ and remains positive definite, and since $\log\det(\cdot)$ is concave on the positive definite cone [101], the composition $\log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$ is concave in $\boldsymbol{w}$.

Thus, $\mathcal{J}(\boldsymbol{w})$ is the sum of a strictly concave term, a concave term, and an affine term, so it is strictly concave on the simplex. The closed simplex is compact and convex, and the ridge term makes the diversity term continuous on it. Hence a global maximizer exists. Strict concavity further implies that this maximizer is unique. □

B.2 Derivation of the Iterative Update Rule

To handle the simplex constraint $\sum w_i = 1$, we introduce a Lagrange multiplier ν. The Lagrangian is:

$$\mathcal{L}(\boldsymbol{w}, \nu) = \sum_{i=1}^{N} w_i c_i + \eta \log|\boldsymbol{\Sigma}_{\boldsymbol{w}}| - \tau \sum_{i=1}^{N} w_i(\log w_i - \log p_i) - \nu\left(\sum_{i=1}^{N} w_i - 1\right). \tag{18}$$

Setting the partial derivative with respect to w_k to zero:

$$\frac{\partial \mathcal{L}}{\partial w_k} = c_k + \eta \frac{\partial}{\partial w_k} \log|\boldsymbol{\Sigma}_{\boldsymbol{w}}| - \tau(\log w_k + 1 - \log p_k) - \nu = 0. \tag{19}$$

Using the derivative of the log-determinant term,

$$\frac{\partial}{\partial w_k} \log|\boldsymbol{\Sigma}_{\boldsymbol{w}}| = \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \frac{\partial \boldsymbol{\Sigma}_{\boldsymbol{w}}}{\partial w_k}\right) = \text{Tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \bar{\boldsymbol{x}}_k \bar{\boldsymbol{x}}_k^\top\right) = \bar{\boldsymbol{x}}_k^\top \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \bar{\boldsymbol{x}}_k = h_k(\boldsymbol{w}), \tag{20}$$

we substitute this back into the stationarity condition:

$$c_k + \eta h_k(\boldsymbol{w}) - \tau \log w_k + \tau \log p_k - \tau - \nu = 0. \tag{21}$$

Rearranging terms to solve for $\log w_k$:

$$\tau \log w_k = \tau \log p_k + c_k + \eta h_k(\boldsymbol{w}) - (\nu + \tau). \tag{22}$$

Exponentiating both sides yields the general solution form:

$$w_k^* \propto p_k \exp\left(\frac{c_k + \eta h_k(\boldsymbol{w})}{\tau}\right). \tag{23}$$

This reveals that the optimal weight is determined by the contribution score, the diversity score, and the prior probability p_k.

Solution under Uniform Prior. In the standard DRIVE setting (as described in Algorithm 1), we assume a non-informative, uniform prior where $p_k = 1/N$ for all samples. Substituting $p_k = 1/N$ into the equation above:

$$w_k^* \propto \frac{1}{N} \exp\left(\frac{c_k + \eta h_k(\boldsymbol{w})}{\tau}\right). \tag{24}$$

Since the constant factor $1/N$ is common to all samples, it cancels out during the normalization step $\sum w_i = 1$. Thus, under the uniform prior assumption, the update rule simplifies to the standard Softmax form:

$$w_k^{(t+1)} = \frac{\exp\left(\frac{c_k + \eta h_k(\boldsymbol{w}^{(t)})}{\tau}\right)}{\sum_{j=1}^{N} \exp\left(\frac{c_j + \eta h_j(\boldsymbol{w}^{(t)})}{\tau}\right)}. \tag{25}$$

This confirms that our standard algorithm is a special case of the general KL-regularized framework where the prior distribution over the data is uniform.

B.3 Optimization for the No-Prior Ablation

The "w/o KL Prior" ablation removes the prior regularization term from Eq. (4). We optimize the remaining contribution–diversity objective

$$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}\in\Delta_{S-1}} \sum_{i=1}^{S} w_i c_i + \eta \log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|, \qquad \boldsymbol{\Sigma}_{\boldsymbol{w}} = \sum_{i=1}^{S} w_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top + \epsilon \boldsymbol{I}.$$

Here, $\Delta_{S-1} = \{\boldsymbol{w} : w_i \geq 0, \sum_i w_i = 1\}$. This objective has no temperature parameter, and its stationarity condition is implicit because $h_i(\boldsymbol{w}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \bar{\boldsymbol{x}}_i$ depends on $\boldsymbol{w}$. For active samples, the KKT condition is $c_i + \eta h_i(\boldsymbol{w}^*) = \nu$, with $c_i + \eta h_i(\boldsymbol{w}^*) \leq \nu$ for inactive samples. Thus, the no-prior objective does not yield a softmax-form closed update.

For consistency with DRIVE, we use the same iterative covariance-refresh framework, but replace the KL-induced softmax step with simplex-constrained ascent. At iteration t, we compute $h_i^{(t)} = \bar{\boldsymbol{x}}_i^\top (\boldsymbol{\Sigma}_{\boldsymbol{w}}^{(t)})^{-1} \bar{\boldsymbol{x}}_i$ and $g_i^{(t)} = c_i + \eta h_i^{(t)}$, then update

$$\boldsymbol{w}^{(t+1)} = \Pi_{\Delta_{S-1}} \left(\boldsymbol{w}^{(t)} + \alpha_t \boldsymbol{g}^{(t)} \right),$$

where $\Pi_{\Delta_{S-1}}$ is Euclidean projection onto the simplex. This keeps weights feasible and uses the same local support set, projected features, ridge-regularized covariance, and iteration budget as full DRIVE.

B.4 Convergence of the Weight Solver

We analyze Eq. (4) with the fixed-point solver in Eq. (5). To match the implementation, we absorb the ridge term into the covariance:

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} = \sum_{i=1}^{N} w_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top + \epsilon \boldsymbol{I}, \qquad \epsilon > 0,$$

and define $h_i(\boldsymbol{w}) = \bar{\boldsymbol{x}}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \bar{\boldsymbol{x}}_i$, with $\boldsymbol{h}(\boldsymbol{w}) = (h_1(\boldsymbol{w}), \ldots, h_N(\boldsymbol{w}))$. Setting $\epsilon = 0$ recovers the unregularized form when the covariance is nonsingular.

Theorem 1 (Uniqueness and Linear Convergence)**.** *Assume $p_i > 0$ for all i, $\tau > 0$, $\eta \geq 0$, and $\epsilon > 0$. Then $J(\boldsymbol{w})$ is τ-strongly concave on Δ_{N-1}, so Eq. (4) admits a unique maximizer $\boldsymbol{w}^*$. Moreover, assume $\boldsymbol{h}$ is L_h-Lipschitz from $(\Delta_{N-1}, \|\cdot\|_1)$ to $(\mathbb{R}^N, \|\cdot\|_\infty)$. If $\eta L_h < \tau$, then $\boldsymbol{w}^*$ is the unique fixed point of Eq. (5), and*

$$\|\boldsymbol{w}^{(t)} - \boldsymbol{w}^*\|_1 \leq \xi^t \|\boldsymbol{w}^{(0)} - \boldsymbol{w}^*\|_1, \qquad \xi := \eta L_h / \tau \in (0, 1).$$

Thus, $\|\boldsymbol{w}^{(T)} - \boldsymbol{w}^*\|_1 = O(\xi^T)$.

Proof. The linear term in $J(\boldsymbol{w})$ is affine, $\log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$ is concave because $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ depends affinely on $\boldsymbol{w}$, and $-\tau \mathrm{KL}(\boldsymbol{w}\|\boldsymbol{p})$ is τ-strongly concave on the simplex. Hence $J(\boldsymbol{w})$ is τ-strongly concave and has a unique maximizer $\boldsymbol{w}^*$.

Let $\Psi(\boldsymbol{w}) = \tau \mathrm{KL}(\boldsymbol{w}\|\boldsymbol{p}) + \iota_{\Delta_{N-1}}(\boldsymbol{w})$. Since Ψ is τ-strongly convex, $\|\nabla\Psi^*(\boldsymbol{a}) - \nabla\Psi^*(\boldsymbol{b})\|_1 \leq \tau^{-1}\|\boldsymbol{a} - \boldsymbol{b}\|_\infty$. Writing the solver map as $T(\boldsymbol{w}) = \nabla\Psi^*(\boldsymbol{c} + \eta \boldsymbol{h}(\boldsymbol{w}))$, for any $\boldsymbol{u}, \boldsymbol{v} \in \Delta_{N-1}$,

$$\|T(\boldsymbol{u}) - T(\boldsymbol{v})\|_1 \leq \tau^{-1}\eta \|\boldsymbol{h}(\boldsymbol{u}) - \boldsymbol{h}(\boldsymbol{v})\|_\infty \leq (\eta L_h/\tau)\|\boldsymbol{u} - \boldsymbol{v}\|_1.$$

Thus T is a contraction when $\xi = \eta L_h/\tau < 1$.

The optimality condition $0 \in -\boldsymbol{c} - \eta \boldsymbol{h}(\boldsymbol{w}^*) + \partial\Psi(\boldsymbol{w}^*)$ is equivalent to $\boldsymbol{w}^* = T(\boldsymbol{w}^*)$. Hence $\boldsymbol{w}^*$ is the unique fixed point, and recursively applying the contraction gives the stated linear rate. Finally, the Lipschitz assumption holds under bounded features: if $\|\bar{\boldsymbol{x}}_i\|_2 \leq R_x$, then $\boldsymbol{\Sigma}_{\boldsymbol{w}} \succeq \epsilon \boldsymbol{I}$ implies $|h_i(\boldsymbol{u}) - h_i(\boldsymbol{v})| \leq R_x^4 \epsilon^{-2} \|\boldsymbol{u} - \boldsymbol{v}\|_1$, so one may take $L_h \leq R_x^4/\epsilon^2$. □

B.5 Justification for Randomized Dimensionality Reduction

To reduce the $O(d^3)$ cost of covariance inversion, DRIVE projects centered features by $\tilde{\boldsymbol{x}}_i = R^\top \bar{\boldsymbol{x}}_i$, where $R \in \mathbb{R}^{d\times k}$ and $k \ll d$, reducing the dominant cost to $O(k^3)$. We use a normalized Gaussian projection, $R_{ab} \sim \mathcal{N}(0, 1/k)$, equivalently $R = G/\sqrt{k}$ with $G_{ab} \sim \mathcal{N}(0,1)$. Since $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ is estimated on an augmented local support of size $S = |\mathcal{B}| + M$ with FIFO buffer $M = 128$, we set $k = 128$ as a practical support-scale-matched sketch dimension. This choice should be interpreted as preserving the dominant low-dimensional geometry, rather than as a full-rank embedding guarantee when $\mathrm{rank}(Z) > k$.

Proposition 2 (Effective subspace embedding for weighted local geometry)**.** *Let $Z = W^{1/2}\bar{X} \in \mathbb{R}^{S\times d}$, where $\bar{X}$ is the centered feature matrix and $W = \mathrm{diag}(w_1, \ldots, w_S)$. Let $U_{\mathrm{eff}} \in \mathbb{R}^{d\times r_{\mathrm{eff}}}$ be an orthonormal basis of an effective r_{eff}-dimensional subspace of $\mathrm{span}(Z^\top)$, with $r_{\mathrm{eff}} \le k$. Let $R \in \mathbb{R}^{d\times k}$ have entries $R_{ab} \sim \mathcal{N}(0, 1/k)$. There exists a universal constant $C > 0$ such that, for any $\varepsilon, \delta \in (0,1)$, if*

$$k \ge C\varepsilon^{-2}\left(r_{\mathrm{eff}} + \log(1/\delta)\right),$$

then with probability at least $1-\delta$,

$$(1-\varepsilon) I_{r_{\mathrm{eff}}} \preceq U_{\mathrm{eff}}^\top R R^\top U_{\mathrm{eff}} \preceq (1+\varepsilon) I_{r_{\mathrm{eff}}}.$$

Equivalently, for every $z \in \mathrm{span}(U_{\mathrm{eff}})$,

$$(1-\varepsilon)\|z\|_2^2 \le \|R^\top z\|_2^2 \le (1+\varepsilon)\|z\|_2^2.$$

Remark. When the full local rank $r = \mathrm{rank}(Z)$ exceeds k, the projected covariance is necessarily rank at most k, so the proposition cannot hold over the entire $\mathrm{span}(Z^\top)$. In this regime, our use of $k = 128$ should be viewed as a low-rank randomized approximation to the dominant local geometry. The approximation quality depends on the effective intrinsic rank of the centered features, and sensitivity to k is evaluated empirically in Appendix F.5.

C Detailed Theoretical Proofs

This appendix supports the theoretical claims in Sec. 4.1. Sec. C.1 proves the linear-regime results on contribution reallocation, adaptation toward $\boldsymbol{\theta}^*$, and diversity-induced conditioning. Sec. C.2 extends the argument to the local nonlinear regime with explicit Hessian remainder control.

C.1 Proof for the Linear Regime in Sec. 4.1

Notation used in this subsection. We work in the linear regression abstraction of the main text, where

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} := \sum_{i=1}^{N} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

denotes the (*uncentered*) weighted Gram matrix. This is the natural object for the linear regression analysis below and matches the main text linear regime. *Important:* in the local nonlinear regime (App. C.2) the symbol $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ is reused, but it instead denotes the *centered* feature covariance $\sum_i w_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^\top$ that appears in the diversity objective. The two usages are intentionally different (mirroring the two regimes in the main text), and we will not mix them within a single subsection.

We justify the two linear-regime claims in Sec. 4.1: (i) unified contribution reallocation improves adaptation along conflicting directions, and (ii) the diversity term $\log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$ improves conditioning, accelerates convergence, and enhances estimation robustness.

Consider the weighted linear regression model $f(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$ with loss

$$\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{N} w_i (y_i - \boldsymbol{\theta}^\top \boldsymbol{x}_i)^2.$$

We inherit $\boldsymbol{\theta}^*$ from the main text as the ideal target parameter, and use $\boldsymbol{\theta}_{\boldsymbol{w}}^{\mathrm{opt}} := \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta})$ to denote the weighted optimum. Define the optimization error $\boldsymbol{\epsilon}(t) := \boldsymbol{\theta}(t) - \boldsymbol{\theta}_{\boldsymbol{w}}^{\mathrm{opt}}$. The gradient flow $\dot{\boldsymbol{\theta}} = -\nabla \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta})$ satisfies

$$\frac{d}{dt}\boldsymbol{\epsilon}(t) = -\boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{\epsilon}(t),$$

so the contraction rate along any unit direction $\boldsymbol{u}$ is governed by the effective eigenvalue $\mu_{\boldsymbol{u}} = \boldsymbol{u}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{u}$.

Contribution-side analysis. Let $\boldsymbol{u}_{\text{con}}$ denote a conflicting direction, i.e., a direction that is important for the target task but mainly supported by samples with high pre-training loss. Under stability-oriented weighting such as FLOW, samples aligned with $\boldsymbol{u}_{\text{con}}$ receive insufficient mass, so

$$\mu_{\text{con}} = \boldsymbol{u}_{\text{con}}^\top \boldsymbol{\Sigma}_{\boldsymbol{w}} \boldsymbol{u}_{\text{con}} = \sum_{i=1}^{N} w_i (\boldsymbol{x}_i^\top \boldsymbol{u}_{\text{con}})^2$$

becomes small, slowing contraction along $\boldsymbol{u}_{\text{con}}$ and possibly stalling adaptation.

DRIVE uses the unified contribution $c_i = \lambda a_i + (1-\lambda) t_i$. Since a_i favors stability while t_i favors task adaptation, decreasing λ increases the relative score of target-relevant conflicting samples. Because the weight update is monotone in c_i, this reallocates mass toward samples that support $\boldsymbol{u}_{\text{con}}$, thereby increasing μ_{con} and accelerating error decay along the conflicting direction.

To connect this to the ideal target $\boldsymbol{\theta}^*$, let $r_i^* := y_i - (\boldsymbol{\theta}^*)^\top \boldsymbol{x}_i$ be the residual of $\boldsymbol{\theta}^*$. Along any unit direction $\boldsymbol{u}$, the projected residual gradient at $\boldsymbol{\theta}^*$ is

$$g_{\boldsymbol{u}}^* \;:=\; \boldsymbol{u}^\top \nabla \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta}^*) = -\sum_i w_i r_i^* (\boldsymbol{x}_i^\top \boldsymbol{u}).$$

Restrict $\boldsymbol{\theta}$ to the one-dimensional slice $\{\boldsymbol{\theta}^* + s\boldsymbol{u} : s \in \mathbb{R}\}$. The loss is quadratic in s:

$$\mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta}^* + s\boldsymbol{u}) = \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta}^*) + s\, g_{\boldsymbol{u}}^* + \tfrac{s^2}{2} \mu_{\boldsymbol{u}},$$

whose minimizer is at $s^* = -g_{\boldsymbol{u}}^* / \mu_{\boldsymbol{u}}$, yielding slice-deviation $|g_{\boldsymbol{u}}^*| / \mu_{\boldsymbol{u}}$. Whenever $\boldsymbol{u}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$, this slice-deviation coincides exactly with the projection of the full deviation $\boldsymbol{\theta}_{\boldsymbol{w}}^{\text{opt}} - \boldsymbol{\theta}^*$ onto $\boldsymbol{u}$. Thus, when the ideal target fits the task-relevant samples well so that $g_{\boldsymbol{u}}^*$ is small, increasing $\mu_{\boldsymbol{u}}$ reduces the projected fitting gap to $\boldsymbol{\theta}^*$. By assigning larger weights to target-relevant conflicting samples, DRIVE can therefore move the model closer to $\boldsymbol{\theta}^*$ along directions that would otherwise be under-weighted by purely stability-oriented weighting. This proves the contribution-side claim.

Diversity-side analysis. Under the standard D-optimal design abstraction, w_i is interpreted as a relative precision or allocation weight, with conditionally independent noise satisfying $\text{Var}(y_i \mid \boldsymbol{x}_i) = \sigma^2 / w_i$. Then the weighted least-squares estimator has covariance $\text{Cov}(\hat{\boldsymbol{\theta}}) = \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}$. More generally, without this precision-weight interpretation, the exact covariance takes a sandwich form. We use the D-optimal abstraction to analyze how enlarging $|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$ reduces generalized variance and improves estimator stability.

To see why it also improves conditioning, by Jacobi's formula

$$\frac{\partial}{\partial w_i} \log |\boldsymbol{\Sigma}_{\boldsymbol{w}}| = \text{tr}\big(\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{x}_i \boldsymbol{x}_i^\top\big) = \boldsymbol{x}_i^\top \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{x}_i.$$

Let $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \sum_{j=1}^{d} \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^\top$ be the eigendecomposition and write $\boldsymbol{x}_i = \sum_j \alpha_{ij} \boldsymbol{u}_j$. Then

$$\frac{\partial}{\partial w_i} \log |\boldsymbol{\Sigma}_{\boldsymbol{w}}| = \sum_{j=1}^{d} \frac{\alpha_{ij}^2}{\lambda_j}.$$

Hence directions with small eigenvalues receive larger gradient pressure. The log-determinant term therefore preferentially lifts under-covered directions, flattens the spectrum, and reduces the condition number $\kappa(\boldsymbol{\Sigma}_{\boldsymbol{w}}) = \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) / \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}})$.

Finally, under discrete-time gradient descent with step size $\eta_{\text{gd}} < 2/\lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{w}})$, the worst-case convergence factor is monotonically improved as $\kappa(\boldsymbol{\Sigma}_{\boldsymbol{w}})$ decreases. Therefore, by enlarging small eigenvalues and improving conditioning, the diversity term accelerates optimization and enhances robustness in the linear regime.

C.2 Local Nonlinear Analysis with Hessian Remainder Control

Notation used in this subsection. We extend the local analysis in Sec. 4.1 by explicitly accounting for the non-Gauss–Newton Hessian component and the local nonlinear remainder. Throughout this subsection, $\boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}) := \nabla^2 \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta})$ is the full weighted Hessian. We use

$$\boldsymbol{G}_{\boldsymbol{w}} \;:=\; \boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{\theta}_{\text{pre}}) \;=\; \sum_{i=1}^{N} w_i \, \boldsymbol{J}_i^{\top} \boldsymbol{B}_i \boldsymbol{J}_i$$

for the weighted Gauss–Newton block evaluated at $\boldsymbol{\theta}_{\text{pre}}$, where $\boldsymbol{G}_{\boldsymbol{w}}(\boldsymbol{\theta})$ is the GN block introduced in the main text, $\boldsymbol{J}_i := \nabla_{\boldsymbol{\theta}} f_i(\boldsymbol{\theta}_{\text{pre}})$ is the per-sample Jacobian of the model output, and $\boldsymbol{B}_i$ is the corresponding output-space loss Hessian. Following the main text local nonlinear regime, we reserve

$$\boldsymbol{\Sigma}_{\boldsymbol{w}} \;:=\; \sum_{i} w_i \bar{\boldsymbol{x}}_i \bar{\boldsymbol{x}}_i^{\top}$$

for the *centered* feature covariance used in the diversity objective $\log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$, where $\bar{\boldsymbol{x}}_i$ is the (mini-batch) centered feature. *Note that this $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ is intentionally distinct from the (uncentered) weighted Gram matrix denoted by the same symbol in App. C.1; the two regimes use the natural object for their respective analyses.* The contribution scores a_i, t_i, the prior p_i, the diversity score h_i, the trade-off λ, the diversity coefficient η and the temperature τ are inherited from the main text.

Consider the local neighborhood $\mathcal{B}_R = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{pre}}\|_2 \le R\}$.

Assumption 1 (Local Hessian smoothness)**.** *There exists $L_H \ge 0$ such that $\|\boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}) - \boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}_{\text{pre}})\|_2 \le L_H \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{pre}}\|_2$ for all $\boldsymbol{\theta} \in \mathcal{B}_R$.*

Assumption 2 (Bounded non-GN residual)**.** *At $\boldsymbol{\theta}_{\text{pre}}$ we have the decomposition $\boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}_{\text{pre}}) = \boldsymbol{G}_{\boldsymbol{w}} + \boldsymbol{E}_{\boldsymbol{w}}$ with $\|\boldsymbol{E}_{\boldsymbol{w}}\|_2 \le \kappa$.*

Assumption 3 (Spectral comparison on the local subspace)**.** *Let $\mathcal{S}$ be the local tangent subspace relevant to the update. There exist $\alpha, \beta > 0$ such that for any unit vector $\boldsymbol{v} \in \mathcal{S}$,*

$$\alpha \, \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) \;\le\; \boldsymbol{v}^{\top} \boldsymbol{G}_{\boldsymbol{w}} \boldsymbol{v} \;\le\; \beta \, \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

Assumption 3 formalizes the link between the diversity-side object $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ (centered feature covariance) and the GN block $\boldsymbol{G}_{\boldsymbol{w}}$, which is the actual driver of local curvature. It is not an a priori statement that the two are equal but rather that they share a comparable spectrum on the directions that the optimizer can move along.

Local strong convexity and smoothness. For any $\boldsymbol{\theta} \in \mathcal{B}_R$, write

$$\boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}) \;=\; \boldsymbol{G}_{\boldsymbol{w}} + \boldsymbol{E}_{\boldsymbol{w}} + \boldsymbol{R}_{\boldsymbol{w}}(\boldsymbol{\theta}), \qquad \boldsymbol{R}_{\boldsymbol{w}}(\boldsymbol{\theta}) \;:=\; \boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}) - \boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}_{\text{pre}}).$$

By Assumption 1, $\|\boldsymbol{R}_{\boldsymbol{w}}(\boldsymbol{\theta})\|_2 \le L_H R$. Combining with Assumptions 2–3 and the triangle inequality, for any unit $\boldsymbol{v} \in \mathcal{S}$ and $\boldsymbol{\theta} \in \mathcal{B}_R$,

$$\alpha \, \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) - \kappa - L_H R \;\le\; \boldsymbol{v}^{\top} \boldsymbol{H}_{\boldsymbol{w}}(\boldsymbol{\theta}) \boldsymbol{v} \;\le\; \beta \, \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) + \kappa + L_H R.$$

Define

$$m_{\boldsymbol{w}} \;:=\; \alpha \, \lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) - \kappa - L_H R, \quad M_{\boldsymbol{w}} \;:=\; \beta \, \lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{w}}) + \kappa + L_H R.$$

If $m_{\boldsymbol{w}} > 0$, then on $\mathcal{S} \cap \mathcal{B}_R$ the weighted objective is locally $m_{\boldsymbol{w}}$-strongly convex and $M_{\boldsymbol{w}}$-smooth, with local condition number bounded by $M_{\boldsymbol{w}}/m_{\boldsymbol{w}}$.

This yields two standard consequences. First, if gradient descent uses step $0 < \zeta \le 1/M_{\boldsymbol{w}}$ and stays in $\mathcal{B}_R \cap \mathcal{S}$, the local objective gap contracts at rate $1 - m_{\boldsymbol{w}}/M_{\boldsymbol{w}}$ per step. Second, if the empirical gradient is perturbed by $\boldsymbol{e}$ with $\|e\|_2 \le \delta$, the corresponding perturbed local stationary point $\tilde{\boldsymbol{\theta}}_{\boldsymbol{w}}$ satisfies

$$\|\tilde{\boldsymbol{\theta}}_{\boldsymbol{w}} - \boldsymbol{\theta}_{\boldsymbol{w}}^{\text{loc}}\|_2 \;\le\; \delta/m_{\boldsymbol{w}},$$

where $\boldsymbol{\theta}_{\boldsymbol{w}}^{\text{loc}}$ denotes the unperturbed local stationary point. Thus, a larger $m_{\boldsymbol{w}}$ improves both local descent and robustness, provided that the nonlinear residual $\kappa + L_H R$ is not dominant.

Connection to the ideal target. Suppose $\boldsymbol{\theta}^* \in \mathcal{B}_R \cap \mathcal{S}$. By local $m_{\boldsymbol{w}}$-strong convexity on $\mathcal{S}$,

$$\|\boldsymbol{\theta}^{\mathrm{loc}}_{\boldsymbol{w}} - \boldsymbol{\theta}^*\|_2 \;\le\; m_{\boldsymbol{w}}^{-1}\, \|\Pi_{\mathcal{S}} \nabla \mathcal{L}_{\boldsymbol{w}}(\boldsymbol{\theta}^*)\|_2,$$

where $\Pi_{\mathcal{S}}$ is the projection onto the local update subspace. Therefore DRIVE can approach the ideal target when (a) the projected weighted residual gradient at $\boldsymbol{\theta}^*$ is small and (b) the effective lower curvature $m_{\boldsymbol{w}}$ is large. The contribution term targets condition (a) by upweighting task-relevant samples that are better explained by $\boldsymbol{\theta}^*$; the diversity term targets condition (b) by improving the centered feature covariance spectrum.

Connecting to the DRIVE weighting rule. With the diversity score fixed, one exponential weighting step satisfies

$$w_i(\lambda) \;\propto\; p_i \exp\big((\lambda a_i + (1-\lambda) t_i + \eta h_i)/\tau\big).$$

For each sample, let $\boldsymbol{J}_i = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}_i; \boldsymbol{\theta}_{\mathrm{pre}})$ denote the model-output Jacobian and let $\boldsymbol{B}_i = \nabla^2_{\boldsymbol{o}} \ell(\boldsymbol{o}, \boldsymbol{y}_i)|_{\boldsymbol{o}=f(\boldsymbol{x}_i;\boldsymbol{\theta}_{\mathrm{pre}})}$ denote the output-space loss Hessian. For a task-relevant direction $\boldsymbol{v} \in \mathcal{S}$, define the per-sample contribution to $\boldsymbol{v}^\top \boldsymbol{G}_{\boldsymbol{w}} \boldsymbol{v}$ as

$$q_i(\boldsymbol{v}) \;:=\; \boldsymbol{v}^\top \boldsymbol{J}_i^\top \boldsymbol{B}_i \boldsymbol{J}_i \boldsymbol{v} \;\ge 0.$$

Differentiating the softmax weights and simplifying yields

$$\frac{d}{d\lambda} \sum_{i=1}^{N} w_i(\lambda)\, q_i(\boldsymbol{v}) \;=\; \tau^{-1}\, \mathrm{Cov}_{w(\lambda)}\big(q_i(\boldsymbol{v}),\, a_i - t_i\big).$$

Therefore, when task-relevant conflicting samples have larger task-adaptation scores, i.e.,

$$\mathrm{Cov}_{w(\lambda)}\big(q_i(\boldsymbol{v}),\, t_i - a_i\big) \;>\; 0,$$

decreasing λ increases the effective spectral mass along $\boldsymbol{v}$, which enlarges $\boldsymbol{v}^\top \boldsymbol{G}_{\boldsymbol{w}} \boldsymbol{v}$ and, through Assumption 3, supports a larger $m_{\boldsymbol{w}}$ and a smaller projected fitting gap toward $\boldsymbol{\theta}^*$.

The diversity term $\log|\boldsymbol{\Sigma}_{\boldsymbol{w}}|$ further discourages collapse of the centered feature covariance: as in App. C.1, its eigenvalue derivative is $1/\lambda_j(\boldsymbol{\Sigma}_{\boldsymbol{w}})$, so it places stronger pressure on under-covered feature directions with small eigenvalues, encouraging a more balanced feature spectrum. Under Assumption 3, this can increase $\lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{w}})$, hence $m_{\boldsymbol{w}}$, and improve the local condition number $M_{\boldsymbol{w}}/m_{\boldsymbol{w}}$ whenever $\kappa + L_H R$ is controlled.

In summary, DRIVE provides a local nonlinear guarantee with explicit Hessian remainder control: contribution reallocation reduces the residual gradient toward the ideal target $\boldsymbol{\theta}^*$ along task-relevant conflicting directions, while diversity regularization improves feature-space coverage and supports better local conditioning. This result is local in $\mathcal{B}_R$ and does not claim a global non-convex guarantee.

D More Experimental Details

In this section, we provide a comprehensive overview of our experimental setup to ensure reproducibility. We begin by detailing the implementation and configurations of the baseline methods in D.1. Subsequently, we provide descriptions of the diverse downstream vision datasets utilized in our experiments in D.2. We then outline the specific evaluation metrics and benchmarks employed for Large Language Models in D.3. Finally, we present the implementation details, including hyperparameter settings and training configurations for both vision and language models, in D.4.

D.1 Baseline Details

We compare DRIVE against SOTA methods spanning parameter-space regularization, PEFT, sample selection, sample reweighting, and token-level loss reweighting. All are implemented under the *data-oblivious* setting (no access to pre-training data).

Standard Fine-Tuning (Standard FT): Updates all parameters using standard cross-entropy loss. It typically suffers severe catastrophic forgetting, serving as the lower bound for general capability preservation.

Linear Probing (LP): Freezes the pre-trained backbone and only trains the task-specific head. While maximizing retention, it often yields sub-optimal task adaptation.

ℓ_2-Regularization [38]: Constrains parameter drift from pre-trained weights θ_{pre} via an ℓ_2 penalty: $\mathcal{L}(\boldsymbol{\theta}) = \sum \ell_i(\boldsymbol{\theta}) + \lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{pre}}\|_2^2$. Higher λ favors retention over adaptation.

WiSE-FT [79]: Linearly interpolates between pre-trained ($\boldsymbol{\theta}_{\text{pre}}$) and fine-tuned ($\boldsymbol{\theta}_{\text{ft}}$) weights: $\boldsymbol{\theta}_{\text{wise}} = (1-\alpha)\boldsymbol{\theta}_{\text{pre}} + \alpha\boldsymbol{\theta}_{\text{ft}}$. We set $\alpha = 0.5$ as pre-training validation data is unavailable.

LoRA [10]: Injects trainable rank decomposition matrices into Transformer layers while freezing the backbone. Limiting trainable parameters inherently mitigates forgetting [11]. We set rank $r = 64$.

DEFT-UCS [22]: A dataset pruning framework that selects a representative "core-set" via K-Means clustering on pre-trained embeddings. We use the recommended "hard sampling" strategy. Unlike DRIVE's continuous weighting, DEFT-UCS performs discrete data sub-sampling prior to training.

FLOW [20]: Dynamically weights samples inversely to their pre-trained loss: $w_i \propto \exp(-\ell_i(\boldsymbol{\theta}_{\text{pre}})/\tau)$, upweighting "easy" samples to anchor the model. FLOW lacks explicit task adaptation mechanisms and ignores diversity.

TALR [6]: Token-Adaptive Loss Reweighting dynamically rescales token-level losses according to predicted token probabilities, down-weighting low-probability hard tokens to reduce their disproportionate impact on general capability degradation.

D.2 Vision Datasets and Evaluation Details

For vision tasks, we primarily utilize Top-1 Accuracy as the evaluation metric. We report the Target Accuracy on the fine-tuning datasets to assess adaptation performance. To monitor catastrophic forgetting, we evaluate the ImageNet-1K Retention Accuracy, where the pre-trained classification head is restored during inference [102]. The Overall Avg is calculated as the arithmetic mean of the Target Accuracy and the ImageNet-1K Retention Accuracy. For the pre-trained loss used to compute a_i, we freeze the ImageNet-pretrained backbone and use a task-specific linear-probe head trained on the downstream training set.

- ImageNet-1K [102]: Large-scale pre-training source with 1,000 diverse classes.
- CIFAR-10 [80]: Standard benchmark with 10 classes of 32×32 images.
- CIFAR-100 [80]: Fine-grained extension of CIFAR-10 with 100 classes.
- Caltech101 [82]: General object recognition benchmark with 101 categories.
- Flowers102 [81]: Fine-grained classification of 102 flower species.
- Cars [83]: Fine-grained dataset containing 196 car models.
- Dogs [84]: Fine-grained dataset covering 120 dog breeds.
- Food101 [85]: Large-scale dataset with 101 food categories.
- Beans [86]: Agricultural dataset for leaf disease classification.
- EuroSAT [87]: Satellite imagery for land-use and land-cover detection.

D.3 LLM Datasets and Evaluation Details

We assess model performance across three dimensions: Commonsense Reasoning, General Capabilities (MMLU, MBPP) to monitor forgetting, and Target Tasks for mathematical reasoning. Code generation capabilities are further evaluated using HumanEval and the Magicoder dataset. For LLM training and contribution scoring, each sequence is treated as one sample, and sample-level cross-entropy is averaged over non-padding target tokens, avoiding length bias.

We employ the lm-evaluation-harness [94] library for standardized assessment. We report Pass@1 for code generation and Accuracy for other tasks. The specific configurations are: General Capabilities: MMLU is evaluated in a 5-shot setting. Commonsense Reasoning: HellaSwag, PIQA, OpenBookQA, ARC-Easy, and ARC-Challenge are evaluated in a 0-shot setting. Mathematical Reasoning: GSM8K is assessed using 5-shot prompting. Code Generation: We evaluate HumanEval in a 0-shot setting and MBPP in a 3-shot setting. For aggregation, we calculate the General Capability score as the average of the three general benchmarks, while the Overall Avg. is computed as the mean of this General Capability score and the Target Task accuracy.

- HellaSwag [103]: A commonsense reasoning benchmark for selecting a plausible continuation.
- ARC (Easy & Challenge) [104]: Grade-school science questions. *ARC-e* targets basic reasoning, while *ARC-c* involves complex questions requiring deep inference.
- PIQA [105]: Evaluates physical commonsense through solutions to everyday physical goals.

- OpenBookQA [106]: Elementary science QA combining open-book facts with world knowledge.
- MMLU [92]: A comprehensive benchmark covering 57 subjects to evaluate knowledge breadth.
- MBPP [93]: Contains 974 basic Python problems requiring code satisfying provided test cases.
- GSM8K [90]: 1,319 grade-school math word problems requiring multi-step arithmetic reasoning.
- MetaMathQA [88]: Bootstraps math questions into instruction data for reasoning.
- Magicoder [89]: Leverages open-source snippets to create coding problems for instruction tuning.
- HumanEval [91]: Evaluates code synthesis via 164 handwritten Python problems with unit tests.

D.4 Implementation Details

D.4.1 Vision Model Implementation Details

For the vision benchmarks, we performed an extensive hyper-parameter search over a grid of learning rates (lr $\in \{5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4\}$) for standard fine-tuning, linear probing, and our proposed method. Experiments were conducted using ResNet18 and ResNet50 architectures across nine downstream datasets: CIFAR-10, CIFAR-100, Flowers102, Caltech101, Stanford Dogs, Stanford Cars, Food101, Beans, and EuroSAT. We use a global batch size of 128 for all vision experiments. We selected the optimal learning rates using a held-out validation split from the downstream training data, and used the final test set only once for reporting.

Table 12: Hyperparameter configurations for DRIVE across various datasets.

Datasets	CIFAR10	CIFAR100	Flowers	Caltech	Dogs	Cars	Food	Beans	EuroSAT
Epochs	20	25	25	30	30	30	30	20	20
ResNet18	1e-3	5e-3	5e-2	1e-2	5e-3	1e-2	1e-2	1e-2	1e-3
ResNet50	5e-4	1e-3	1e-2	1e-2	5e-3	1e-2	1e-2	1e-2	1e-3

To ensure a fair comparison, we maintained consistent training configurations across all baselines. Specifically, we utilized an SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The learning rate followed a Cosine Annealing schedule for all fine-tuning tasks, whereas linear probing employed a constant learning rate.

D.4.2 LLMs Implementation Details

All Large Language Models (LLMs) are fine-tuned using the AdamW optimizer paired with a cosine learning rate scheduler. To ensure a fair comparison, we enforce a uniform training configuration across all models (Gemma-2-2B, Llama-3.2-3B, and Llama-3.1-8B). Regarding dataset-specific configurations, for MetaMathQA, we set the global batch size to 128 with a maximum sequence length of 1024; for Magicoder, we use a global batch size of 64 and a maximum sequence length of 4096. Common settings include a warmup ratio of 0.03 and a training duration of 3 epochs. To ensure statistical reliability, all results are reported as the mean and standard deviation across five independent runs. All experiments were conducted on four NVIDIA A100 GPUs. For LoRA baselines, the rank is set to $r = 64$. The learning rate is tuned for each method–model combination using validation performance. LoRA uses 2e−4 following the PEFT setting. All full-parameter methods, including Standard FT, FLOW, ℓ_2-Reg., DEFT-UCS, and DRIVE, share the same grid {5e−6, 1e−5, 2e−5}. DRIVE selects 2e−5 for Gemma-2-2B and Llama-3.2-3B, and 1e−5 for Llama-3.1-8B.

E Additional Experimental Results and Analysis

In this appendix, we provide a comprehensive empirical analysis to complement the main text. We begin by detailing fine-grained performance across diverse downstream vision datasets in Appendix E.1 and verifying architectural generalization on Vision Transformers in Appendix E.2. Subsequently, Appendix E.3 extends the LLM evaluation to code generation tasks. We then validate our design choices through in-depth sensitivity analyses: we examine the magnitude and temporal schedule of the balance parameter λ in Appendix E.4 and Appendix E.5, respectively, followed by the prior regularization threshold τ in Appendix E.6, the diversity coefficient η in Appendix E.7, and component-wise contributions on vision benchmarks in Appendix E.8. Finally, we report the computational efficiency of DRIVE in Appendix E.9, demonstrating its scalability compared to baselines.

Table 13: Combined performance comparison across six datasets. For each dataset, **Tgt** denotes the target task accuracy (higher is better for adaptation), and **Gen** denotes the ImageNet-1K accuracy (higher is better for anti-forgetting) (Accuracy in %). Best results are **bold**, second-best results are underlined.

	Method	Caltech101		Dogs		Cars		Food101		Beans		EuroSAT	
		Gen	Tgt	Gen	Tgt	Gen	Tgt	Gen	Tgt	Gen	Tgt	Gen	Tgt
ResNet18	Linear probing	**69.76**	88.63	**69.76**	78.55	**69.76**	50.21	**69.76**	60.03	**69.76**	85.78	**69.76**	91.35
	Standard FT	35.79	**94.89**	58.01	80.30	1.84	**87.64**	0.23	74.03	48.95	**97.66**	24.79	96.79
	ℓ_2-Regularization	45.16	**94.89**	60.11	81.33	20.79	83.79	10.20	75.69	40.77	96.88	39.09	95.42
	WiSE-FT	62.96	90.60	66.89	75.42	32.57	65.46	20.79	70.30	44.87	96.41	50.75	90.18
	DEFT-UCS	59.69	89.59	59.98	79.62	28.54	60.31	10.77	70.94	41.97	95.31	50.32	90.01
	FLOW	65.72	90.46	66.31	82.22	57.65	61.85	53.79	73.43	65.79	94.06	66.00	96.26
	DRIVE (Ours)	64.91	93.90	66.95	**84.82**	55.89	76.72	53.61	**77.48**	64.78	94.84	65.78	**97.17**
ResNet50	Linear probing	**80.10**	92.04	**80.10**	85.40	**80.10**	42.78	**80.10**	53.08	**80.10**	86.41	**80.10**	92.35
	Standard FT	42.53	**97.73**	69.25	87.15	0.92	**90.27**	0.42	76.80	53.79	**98.75**	34.79	97.04
	ℓ_2-Regularization	45.73	95.82	78.96	82.86	14.61	89.97	17.63	76.11	60.21	95.31	66.46	94.21
	WiSE-FT	74.76	92.71	74.56	74.30	10.84	63.79	27.87	72.33	70.64	92.34	71.54	95.43
	DEFT-UCS	60.78	90.68	72.54	80.67	26.87	67.15	22.95	71.23	64.56	90.47	67.64	92.33
	FLOW	71.63	92.79	75.57	89.63	67.98	70.76	55.78	77.68	77.19	95.00	75.32	96.37
	DRIVE (Ours)	71.03	94.96	73.04	**89.94**	66.91	77.68	67.40	**79.37**	76.34	96.41	74.55	**98.44**

E.1 Analysis of Vision Tasks

To evaluate the robustness of DRIVE, we present the fine-grained performance on six diverse downstream vision datasets: Caltech101, Stanford Dogs, Stanford Cars, Food101, Beans, and EuroSAT.

Target Task Adaptation. Table 13 details the top-1 accuracy on each specific target dataset after fine-tuning. DRIVE consistently demonstrates strong performance across both ResNet-18 and ResNet-50 architectures. Notably, on fine-grained classification tasks such as Stanford Cars and Food101, DRIVE achieves significant improvements over competing baselines. For instance, on the Food101 dataset with ResNet-50, DRIVE outperforms FLOW, an anti-forgetting oriented weighting method, while maintaining a superior adaptation score compared to other methods.

Anti-forgetting Capability. Table 13 reports the Top-1 ImageNet-1K accuracy (the pre-training task) after fine-tuning. A major challenge in fine-tuning is the catastrophic drop in pre-trained performance, which is particularly severe in Standard FT. DRIVE effectively mitigates this issue, maintaining significantly higher retention accuracy compared to Standard FT. Furthermore, DRIVE demonstrates anti-forgetting capabilities that are competitive with or superior to ℓ_2-Regularization, while simultaneously achieving higher accuracy on the target tasks.

Consolidated Performance Analysis. Table 14 presents the average accuracy across the six datasets, reflecting the overall balance between representation preservation and task adaptation. Across both ResNet-18 and ResNet-50 architectures, DRIVE achieves the highest average accuracy on the majority of datasets, including Food101 and EuroSAT. While methods like Linear Probing preserve pre-trained knowledge perfectly, they often fail to adapt effectively to new tasks, resulting in lower overall average scores compared to DRIVE. This demonstrates that DRIVE achieves a robust stability-plasticity trade-off, successfully resolving the dilemma.

E.2 Experiments on Vision Transformers

To verify the architectural generalization of DRIVE beyond Convolutional Neural Networks (CNNs), we extended our evaluation to the Vision Transformer (ViT) architecture. Specifically, we fine-tuned a pre-trained ViT-B/16[107] model on the Food101 dataset, a challenging fine-grained classification benchmark.

As presented in Table 15, the results verify that DRIVE generalizes effectively to Transformer-based architectures, offering a superior trade-off compared to existing baselines on the challenging fine-grained Food101 benchmark.

While Standard FT achieves high target adaptation, it suffers from severe catastrophic forgetting. Specifically, the ImageNet-1K accuracy drops by over 24% compared to the pre-trained baseline, confirming that the high plasticity of ViTs leads to significant drift from the general manifold without regularization.

Table 14: Comparison of Average Accuracy (%) across six downstream vision datasets. This metric reflects the overall balance between representation preservation and task adaptation. Best results are **bold**, second-best results are underlined.

Model	Method	Caltech101	Dogs	Cars	Food101	Beans	EuroSAT
ResNet18	Linear probing	79.20	74.16	59.99	64.90	77.77	80.56
	Standard FT	65.34	69.16	44.74	37.13	73.31	60.79
	ℓ_2-Regularization	70.03	70.72	52.29	42.95	68.83	67.26
	WiSE-FT	76.78	71.16	49.02	45.55	70.64	70.47
	DEFT-UCS	74.64	69.80	44.43	40.86	68.64	70.17
	FLOW	78.09	74.27	59.75	63.61	**79.93**	81.13
	DRIVE (Ours)	**79.41**	**75.89**	**66.31**	**65.55**	79.81	**81.48**
ResNet50	Linear probing	**86.07**	**82.75**	61.44	66.59	83.26	86.23
	Standard FT	70.13	78.20	45.60	38.61	76.27	65.92
	ℓ_2-Regularization	70.78	80.91	52.29	46.87	77.76	80.34
	WiSE-FT	83.74	74.43	37.32	50.10	81.49	83.49
	DEFT-UCS	75.73	76.61	47.01	47.09	77.52	79.99
	FLOW	82.21	82.60	69.37	66.73	86.10	85.85
	DRIVE (Ours)	83.00	81.49	**72.30**	**73.39**	**86.38**	**86.50**

DRIVE successfully mitigates this issue, recovering the general accuracy by nearly 18% compared to Standard FT. Crucially, this substantial gain in stability comes with negligible cost to adaptation, as DRIVE maintains a competitive target accuracy comparable to full fine-tuning. DRIVE demonstrates better adaptability than rigid baselines. It surpasses Linear Probe by more than 7.5% on the target task, proving that DRIVE effectively learns task-specific features where simple probing fails. Furthermore, DRIVE outperforms the baseline FLOW by approximately 1.6% in target accuracy, suggesting that our contribution-based weighting captures fine-grained patterns more effectively than loss-based strategies.

Table 15: Performance comparison (Accuracy in %) of ViT-B/16 on Food101. Best in **bold**, second in underlined.

Method	IN-1K	Food101	Avg.
Pre-trained	82.95	-	-
Standard FT	58.42	**92.24**	75.33
Linear Probe	**82.95**	83.93	83.44
l_2-Reg.	57.94	91.78	74.86
WiSE-FT	55.38	90.64	73.01
DEFT-UCS	67.35	90.15	78.75
FLOW	76.72	89.86	83.29
DRIVE (Ours)	76.15	91.45	**83.80**

E.3 Experiments on Language Models

Table 16 presents results of fine-tuning LLMs on Magicoder. We evaluate target-task performance using HumanEval and general capability retention using Commonsense Reasoning (CS), MMLU, and GSM8K. DRIVE consistently achieves the highest overall average across evaluated models, demonstrating a strong balance between stability and plasticity. Particularly on Llama-3.2-3B, DRIVE improves the overall score by 2.86% over LoRA and by 0.98% over FLOW, while substantially improving general capability retention over Standard FT. Although some retention-oriented baselines, such as TALR on Llama-3.2-3B, obtain slightly higher general capability, they achieve lower HumanEval accuracy. Overall, DRIVE attains the best overall score across all evaluated models.

E.4 Impact of Balance Parameter λ on Language Tasks

The parameter $\lambda \in [0, 1]$ controls the trade-off between anti-forgetting and task adaptation. Table 17 illustrates this spectrum on Llama-3.2-3B.

- **Task Adaptation Focus ($\lambda = 0.2$):** When λ is low, the model prioritizes the target task, achieving a high GSM8K score. However, this comes at the cost of general capabilities, with significant drops in Commonsense (CS) and MMLU scores.
- **Anti-forgetting Focus ($\lambda = 0.8$):** Conversely, a high λ preserves general capabilities but restricts the model's plasticity, leading to lower performance on the target task.
- **Optimal Balance:** The dynamic schedule used in DRIVE (decaying λ) or a balanced fixed value effectively navigates this trade-off, achieving the best overall averaged performance.

Table 16: Performance comparison of models fine-tuned on the Magicoder dataset across various benchmarks. Best results are **bolded** and second-best results are underlined.

Model	Method	General Capability Acc. (%)				Target Acc. (%)	Overall (%)
		CS	MMLU	GSM8K	Avg	HumanEval	Avg
Gemma-2-2B	Pre-trained	$57.23_{\pm 0.00}$	$53.14_{\pm 0.00}$	$23.65_{\pm 0.00}$	$44.67_{\pm 0.00}$	$20.12_{\pm 0.00}$	$32.40_{\pm 0.00}$
	Standard FT	$55.71_{\pm 1.17}$	$50.64_{\pm 0.41}$	$20.96_{\pm 2.04}$	$42.44_{\pm 0.94}$	**35.49**$_{\pm 0.51}$	$38.97_{\pm 0.62}$
	WiSE-FT	$56.86_{\pm 1.59}$	$49.50_{\pm 0.39}$	$22.87_{\pm 2.18}$	$43.08_{\pm 0.86}$	$30.73_{\pm 0.70}$	$36.91_{\pm 0.34}$
	ℓ_2-Reg	$54.72_{\pm 1.35}$	$49.45_{\pm 0.40}$	$20.64_{\pm 2.15}$	$41.60_{\pm 1.16}$	$30.49_{\pm 0.43}$	$36.05_{\pm 0.49}$
	LoRA	$56.27_{\pm 1.65}$	$51.74_{\pm 0.40}$	$21.49_{\pm 2.87}$	$43.17_{\pm 1.26}$	$32.32_{\pm 0.61}$	$37.75_{\pm 0.69}$
	DEFT-UCS	$51.46_{\pm 1.35}$	$50.65_{\pm 0.40}$	$20.65_{\pm 2.46}$	$40.92_{\pm 1.13}$	$33.05_{\pm 0.51}$	$36.99_{\pm 0.57}$
	FLOW	$57.84_{\pm 1.22}$	**53.05**$_{\pm 0.42}$	$23.91_{\pm 2.09}$	$44.93_{\pm 0.96}$	$33.66_{\pm 0.67}$	$39.30_{\pm 0.55}$
	TALR	$58.02_{\pm 1.20}$	$52.92_{\pm 0.41}$	$24.37_{\pm 2.08}$	$45.10_{\pm 0.97}$	$33.54_{\pm 0.43}$	$39.32_{\pm 0.35}$
	DRIVE	**58.24**$_{\pm 1.18}$	$52.56_{\pm 0.40}$	**24.94**$_{\pm 2.11}$	**45.25**$_{\pm 0.28}$	$35.24_{\pm 0.51}$	**40.25**$_{\pm 0.37}$
Llama-3.2-3B	Pre-trained	$54.62_{\pm 0.00}$	$56.54_{\pm 0.00}$	$27.75_{\pm 0.00}$	$46.30_{\pm 0.00}$	$28.66_{\pm 0.00}$	$37.48_{\pm 0.00}$
	Standard FT	$52.46_{\pm 1.17}$	$52.18_{\pm 0.39}$	$25.37_{\pm 2.19}$	$43.34_{\pm 0.90}$	**43.41**$_{\pm 0.67}$	$43.38_{\pm 0.62}$
	WiSE-FT	$50.79_{\pm 1.69}$	$52.84_{\pm 0.40}$	$25.67_{\pm 2.03}$	$43.10_{\pm 0.86}$	$37.80_{\pm 0.43}$	$40.45_{\pm 0.42}$
	ℓ_2-Reg	$54.58_{\pm 1.32}$	$52.94_{\pm 0.41}$	**28.57**$_{\pm 2.21}$	$45.36_{\pm 0.45}$	$40.12_{\pm 0.51}$	$42.74_{\pm 0.26}$
	LoRA	$52.78_{\pm 1.18}$	$53.79_{\pm 0.40}$	$24.09_{\pm 1.33}$	$43.55_{\pm 0.55}$	$40.12_{\pm 0.51}$	$41.84_{\pm 0.47}$
	DEFT-UCS	$50.75_{\pm 1.28}$	$48.15_{\pm 0.42}$	$22.14_{\pm 1.98}$	$40.35_{\pm 0.81}$	$41.46_{\pm 0.61}$	$40.91_{\pm 0.64}$
	FLOW	$56.59_{\pm 1.18}$	$55.91_{\pm 0.39}$	$25.78_{\pm 2.19}$	$46.09_{\pm 0.88}$	$41.34_{\pm 0.51}$	$43.72_{\pm 0.33}$
	TALR	**56.72**$_{\pm 1.17}$	**55.96**$_{\pm 0.40}$	$26.22_{\pm 2.10}$	**46.30**$_{\pm 0.80}$	$41.22_{\pm 0.55}$	$43.76_{\pm 0.52}$
	DRIVE	$55.79_{\pm 1.21}$	$55.49_{\pm 0.40}$	$26.66_{\pm 2.08}$	$45.98_{\pm 0.65}$	**43.41**$_{\pm 0.51}$	**44.70**$_{\pm 0.39}$
Llama-3.1-8B	Pre-trained	$61.59_{\pm 0.00}$	$65.32_{\pm 0.00}$	$50.95_{\pm 0.00}$	$59.29_{\pm 0.00}$	$34.15_{\pm 0.00}$	$46.72_{\pm 0.00}$
	Standard FT	$52.02_{\pm 1.25}$	$57.28_{\pm 0.41}$	$40.41_{\pm 2.05}$	$49.90_{\pm 0.88}$	**65.49**$_{\pm 0.55}$	$57.70_{\pm 0.50}$
	WiSE-FT	$55.79_{\pm 1.45}$	$60.52_{\pm 0.40}$	$43.68_{\pm 1.95}$	$53.33_{\pm 0.73}$	$60.73_{\pm 0.55}$	$57.03_{\pm 0.49}$
	ℓ_2-Reg	$56.69_{\pm 1.38}$	$62.17_{\pm 0.39}$	$43.79_{\pm 2.10}$	$54.22_{\pm 0.92}$	$60.61_{\pm 0.55}$	$57.42_{\pm 0.63}$
	LoRA	$55.44_{\pm 1.22}$	$59.33_{\pm 0.41}$	$45.44_{\pm 1.88}$	$53.40_{\pm 0.58}$	$61.22_{\pm 0.33}$	$57.31_{\pm 0.39}$
	DEFT-UCS	$57.58_{\pm 1.30}$	$62.83_{\pm 0.40}$	$42.17_{\pm 1.98}$	$54.19_{\pm 0.87}$	$57.20_{\pm 0.67}$	$55.70_{\pm 0.19}$
	FLOW	$59.24_{\pm 1.19}$	$63.29_{\pm 0.39}$	$49.64_{\pm 2.15}$	$57.39_{\pm 0.97}$	$62.80_{\pm 0.43}$	$60.10_{\pm 0.50}$
	TALR	$59.58_{\pm 1.20}$	$63.34_{\pm 0.39}$	**49.73**$_{\pm 2.10}$	$57.55_{\pm 0.77}$	$62.68_{\pm 0.67}$	$60.12_{\pm 0.63}$
	DRIVE	**60.51**$_{\pm 1.32}$	**63.46**$_{\pm 0.40}$	$49.51_{\pm 2.02}$	**57.83**$_{\pm 0.59}$	$64.27_{\pm 0.33}$	**61.05**$_{\pm 0.34}$

Table 17: Comparison of different λ on Llama-3.2-3B. Higher λ favors general capabilities, while lower λ benefits target adaptation. $\lambda = 0.8$ achieves the best overall balance.

λ value	General Capability Acc. (%)				Target Acc. (%)	Overall (%)
	CS	MMLU	MBPP	Avg	GSM8K	Avg
$\lambda = 0.2$	42.23	55.55	39.20	45.66	**63.81**	54.74
$\lambda = 0.5$	54.95	55.64	40.40	50.33	60.82	55.58
$\lambda = 0.8$	**55.79**	**55.86**	**41.64**	**51.10**	60.29	**55.70**

E.5 Temporal Dynamics of λ: Increasing vs. Decaying Schedule

To validate the temporal importance of the trade-off parameter, we conduct an ablation study using a linearly increasing schedule for λ ($0 \rightarrow 1$). This configuration prioritizes task adaptation during the initial phase of training, serving as a counter-factual baseline to our proposed decay strategy ($1 \rightarrow 0$). The results across three representative LLMs are summarized in Table 18.

Compared to the standard DRIVE results (see Table 3), this reverse schedule yields consistently sub-optimal outcomes. For instance, the Overall Avg score for Gemma-2-2B drops to 54.92%, while Llama-3.1-8B achieves only 66.75%. These findings underscore that "initial anchoring" to the pre-trained distribution is indispensable. In the early stages of fine-tuning, when gradient updates are typically largest, a low λ fails to provide sufficient regularization (via the KL-divergence term) against the prior distribution. This allows the model to deviate prematurely from the general knowledge manifold, a distortion that subsequent increases in λ cannot fully rectify. Consequently, maintaining a high λ at the onset is crucial for resolving the stability-plasticity dilemma.

E.6 Impact of Prior Regularization τ

The hyperparameter τ controls weight smoothing. In our implementation, τ is set by the percentile of pre-training loss for scale invariance. Table 19 evaluates different percentiles on Llama-3.2-3B. The 10th percentile gives the best overall performance in this LLM setting. A sharper distribution better preserves the contribution-diversity signal while still assigning non-zero weights to other samples. Larger τ smooths weights toward the prior. With a uniform prior, overly large τ makes the weights closer to uniform, weakens DRIVE's weighting signal, and can approach Standard FT-like behavior, degrading the overall trade-off.

Table 18: Performance on GSM8K and general capabilities using a linearly increasing schedule for λ. All accuracies are reported in %.

Model	General Capability Acc. (%)				Target Acc. (%)	Overall (%)
	CS	MMLU	MBPP	Avg	GSM8K	Avg
Gemma-2-2B	56.57	52.59	29.00	46.05	63.79	54.92
Llama-3.2-3B	54.28	54.23	37.00	48.50	63.12	55.81
Llama-3.1-8B	58.48	62.29	45.80	55.52	77.98	66.75

Table 19: Comparison of different τ percentiles on Llama-3.2-3B. All accuracies are reported in %.

τ-Percentile(%)	General Capability Acc. (%)				Target Acc. (%)	Overall (%)
	CS	MMLU	MBPP	Avg	GSM8K	Avg
10	**55.92**	55.71	**41.00**	**50.88**	**64.50**	**57.69**
30	55.19	55.80	38.80	49.93	61.90	55.92
50	54.91	54.34	38.80	49.35	60.80	55.08
70	55.06	55.65	39.60	50.10	59.14	54.62
90	53.13	**55.90**	40.00	49.68	58.01	53.85

E.7 Impact of Diversity Regularization η on Language Tasks

The diversity coefficient η controls the strength of the covariance determinant term. Table 20 studies the sensitivity of DRIVE to this coefficient on language tasks.

- **Effect of Diversity Strength:** General capability metrics vary non-monotonically as η increases. Moderate diversity regularization improves the general capability average, peaking at $\eta = 0.75$, while overly large values reduce CS, MMLU, and MBPP. This suggests diversity regularization is beneficial when balanced, but excessive emphasis can hurt capability preservation.
- **Robustness:** DRIVE remains relatively robust across a range of η values, with the best overall performance observed at $\eta = 1.00$. This setting achieves the highest GSM8K accuracy and overall average, indicating that a moderate-to-strong diversity coefficient provides the best balance between target adaptation and general capability retention.

Table 20: Comparison of different η on Llama-3.2-3B.

η value	General Capability Acc. (%)				Target Acc. (%)	Overall (%)
	CS	MMLU	MBPP	Avg	GSM8K	Avg
0.25	55.08	56.30	40.88	50.75	59.67	55.21
0.50	54.56	56.67	**41.52**	50.92	58.65	54.79
0.75	55.67	**56.79**	41.00	**51.15**	60.61	55.88
1.00	**55.92**	55.71	41.00	50.88	**64.50**	**57.69**
1.25	55.19	55.80	38.80	49.93	61.90	55.92
1.5	54.26	54.21	37.00	48.49	61.50	55.00

E.8 Ablation Study on Vision Tasks

To verify the generalizability of our findings beyond LLMs, we conducted an ablation study on the ResNet-50 vision model (Table 21). The results indicate that removing any individual component of DRIVE compromises the overall performance, with the full method consistently achieving the highest average accuracy across both benchmarks (72.34% on CIFAR-100 and 83.85% on Flowers102). Specifically, the diversity regularization ($\eta = 0$) proves critical for adaptation; its removal leads to a substantial decline in target accuracy, such as a 7.27% drop on CIFAR-100 and a 3.74% decrease on Flowers102. This confirms that expanding the feature volume is necessary to capture task-specific nuances effectively. Additionally, the uniform-prior KL regularization remains the foundation of the framework, as its removal causes the most significant performance degradation, reducing the average accuracy on CIFAR-100 by approximately 9.5% due to weight collapse.

Furthermore, the method navigates the stability-plasticity dilemma through the balance parameter. While the setting without anti-forgetting ($\lambda = 0$) yields the highest target accuracy (86.40% on CIFAR-100 and 93.67% on Flowers102), it incurs a retention cost, reducing ImageNet performance to 57.49% and 71.98%, respectively. Conversely, DRIVE maintains equilibrium. It secures competitive adaptation performance (83.03% on CIFAR-100 and 92.67% on Flowers102) while retaining more knowledge than the plasticity-only baseline. This demonstrates that DRIVE optimizes the joint trade-off, ensuring reliable adaptation without erasing pre-trained representations.

Table 21: Ablation study of DRIVE components on ResNet-50 (Accuracy in %).

Method	CIFAR-100			Flowers102		
	IN-1K	Target Acc.	Avg.	IN-1K	Target Acc.	Avg.
DRIVE w/o Diversity ($\eta = 0$)	**65.79**	75.76	70.78	**77.93**	88.93	83.43
DRIVE w/o KL Prior	55.93	69.71	62.82	66.94	87.52	77.23
DRIVE w/o anti-forgetting ($\lambda = 0$)	57.49	**86.40**	71.95	71.98	**93.67**	82.83
DRIVE w/o task-adaptation ($\lambda = 1$)	63.49	80.91	72.20	76.26	90.73	83.50
DRIVE	61.64	83.03	**72.34**	75.02	92.67	**83.85**

E.9 Computational Efficiency Analysis

We first evaluate the computational overhead on vision tasks using CIFAR-100 with ResNet-18 and ResNet-50 on a single NVIDIA A100 GPU. As shown in Table 22, we measure the average GPU time per epoch. Compared to Standard FT, FLOW introduces slight latency (3.3%–4.6%). DRIVE incurs a moderate increase (11.8%–14.7%) due to projection and determinant calculations required for diversity regularization, yet remains efficient for standard backbones.

Table 22: Training efficiency on CIFAR-100, reported as GPU seconds per epoch.

Model	Method	GPU Time / Epoch (s)	Overhead (%)
ResNet-18	Standard FT	280	-
	FLOW	293	+4.64%
	DRIVE (Ours)	313	+11.79%
ResNet-50	Standard FT	389	-
	FLOW	402	+3.34%
	DRIVE (Ours)	446	+14.65%

We further evaluate scalability on MetaMathQA using Gemma-2B (single A100) and Llama-3-8B (dual A100s). Table 23 reports the average GPU time per epoch. Both methods incur manageable costs: FLOW adds approximately 10% overhead for computing loss distributions, while DRIVE adds roughly 17% due to simultaneous forward passes and covariance computation. Notably, the Johnson-Lindenstrauss transform ($d = 4096 \rightarrow k = 128$) ensures that high-dimensional geometric calculations remain linear, preventing prohibitive scaling on larger models.

Table 23: Training efficiency on MetaMathQA, reported as GPU hours per epoch.

Model	Method	GPU Time / Epoch (h)	Overhead (%)
Gemma-2-2B	Standard FT	5.45	-
	FLOW	6.00	+10.09%
	DRIVE (Ours)	6.38	+17.06%
Llama-3.1-8B	Standard FT	19.42	-
	FLOW	21.36	+9.99%
	DRIVE (Ours)	22.75	+17.15%

E.10 Paired Statistical Analysis of the Main LLM Results

We assess the statistical reliability of the Overall improvements in Table 3 using original training-log scores from five runs with matched random seeds and data orders. DRIVE is paired with the strongest baseline for each model: TALR for Gemma-2-2B and Llama-3.2-3B, and LoRA for Llama-3.1-8B. No additional training or reruns are performed.

For each matched seed, we compute DRIVE's Overall improvement and apply a two-sided paired t-test against zero mean improvement. Displayed Overall scores are calculated from rounded com-

Table 24: Paired statistical analysis of DRIVE against the strongest baseline for each LLM, using five runs with matched random seeds and data orders. Overall scores are percentages; differences and confidence intervals are expressed in percentage points. Positive Δ favors DRIVE.

Model	Baseline	Baseline Overall	DRIVE Overall	Δ	95% CI	Raw p	Holm-adjusted p	Hedges' g_z
Gemma-2-2B	TALR	53.87	55.55	+1.68	[0.85, 2.51]	0.0049	0.0148	2.00
Llama-3.2-3B	TALR	56.26	57.69	+1.43	[0.46, 2.42]	0.0152	0.0152	1.45
Llama-3.1-8B	LoRA	65.70	67.27	+1.57	[0.77, 2.37]	0.0055	0.0148	1.95

ponent means, and Δ is their difference. The 95% confidence intervals, raw p-values, and small-sample-corrected paired effect sizes Hedges' g_z are computed from the original unrounded paired scores. Table 24 reports these statistics and Holm-adjusted p-values across the three model-wise tests; confidence intervals are reported for individual comparisons without multiplicity adjustment.

All three paired comparisons favor DRIVE, with individual 95% confidence intervals excluding zero. All improvements remain significant after Holm correction (adjusted $p \leq 0.0152$). Hedges' g_z ranges from 1.45 to 2.00, indicating large improvements relative to the variability of paired differences. These results support the statistical reliability of DRIVE's reported 1.43–1.68 percentage-point Overall gains over the strongest evaluated baselines under this five-seed protocol.

F Analysis of Contribution Assessment and Diversity Mechanisms

We analyze DRIVE's contribution assessment, diversity modeling, and covariance approximation. Sec. F.1 studies sample-level stability–plasticity conflict via Shapley-style analysis. Sec. F.2 discusses diversity metrics and DRIVE-Cluster. Sec. F.3 compares covariance approximations. Sec. F.4 examines hard-sample adaptation dynamics; Sec. F.5 reports sensitivity to projection dimension k.

F.1 Shapley Analysis of Anti-Forgetting vs. Task Adaptation

To validate the *Stability-Plasticity Dilemma* at the sample level and justify our contribution-based weighting, we conducted an analysis using the Shapley value framework. We utilize the Gradient Inner Product (GIP) [71, 108] as an efficient approximation of the One-Run Shapley value to quantify the contribution of each training sample toward two goals: (1) preserving pre-trained knowledge (Anti-forgetting, a_i) and (2) learning the new task (Task Adaptation, t_i).

F.1.1 Experimental Setup

Full-Batch Validation Gradient Calculation. We use full validation sets at every training step to ensure accurate estimation of the optimization direction. Let $\mathcal{D}_{\text{val}}^{\text{target}}$ and $\mathcal{D}_{\text{val}}^{\text{retain}}$ denote the complete validation sets for the target task (e.g., Flowers102) and the retention task (e.g., ImageNet subset), respectively. At each training step t, we compute the exact average gradient over these datasets:

$$\mathbf{g}_t^{\text{target}} := \nabla_\theta \left[\frac{1}{|\mathcal{D}_{\text{val}}^{\text{target}}|} \sum_{(x,y)\in\mathcal{D}_{\text{val}}^{\text{target}}} \ell\big(f(x;\theta_t), y\big) \right], \tag{26}$$

$$\mathbf{g}_t^{\text{retain}} := \nabla_\theta \left[\frac{1}{|\mathcal{D}_{\text{val}}^{\text{retain}}|} \sum_{(x,y)\in\mathcal{D}_{\text{val}}^{\text{retain}}} \ell_{\text{retain}}\big(f(x;\theta_t), y\big) \right]. \tag{27}$$

Cumulative Sample Contribution. Following [71], we quantify sample utility by summing marginal contributions along the optimization trajectory. Let $\mathcal{T}_i$ be steps where sample x_i appears in the mini-batch. We define Shapley-style anti-forgetting (a_i^{shap}) and task-adaptation (t_i^{shap}) contributions as cumulative alignment between the sample gradient and full-batch validation gradients:

$$a_i^{\text{shap}} := \sum_{t\in\mathcal{T}_i} \underbrace{\eta_t \left\langle \nabla_\theta \ell(f(x_i;\theta_t), y_i),\ \mathbf{g}_t^{\text{retain}} \right\rangle}_{\text{Step-wise Contribution to Retention}}, \qquad t_i^{\text{shap}} := \sum_{t\in\mathcal{T}_i} \underbrace{\eta_t \left\langle \nabla_\theta \ell(f(x_i;\theta_t), y_i),\ \mathbf{g}_t^{\text{target}} \right\rangle}_{\text{Step-wise Contribution to Adaptation}}. \tag{28}$$

Here, η_t denotes the learning rate [71]. The inner product measures geometric alignment, where a positive value indicates the sample facilitates loss minimization on the corresponding validation set.

F.1.2 Empirical Results and Analysis

Validation of Loss-based Proxies. In the main text, we proposed loss-based metrics to define the contribution scores (e.g., $a_i \propto -\ell_{\text{pre}}$, $t_i \propto \ell_{\text{cur}}$). To verify these heuristic proxies, we analyzed their correlation with theoretically grounded Shapley values calculated via Eq. (28). Our empirical analysis reveals a significant positive correlation. The loss-based anti-forgetting metric a_i strongly correlates with the reference a_i^{shap} (Pearson's $r = 0.6493$), corroborating that samples with lower pre-training loss generally contribute disproportionately to parameter stability. The strong positive correlation between t_i and t_i^{shap} ($r = 0.8679$) confirms high-loss samples drive adaptation. This validates using our efficient loss-based proxy in DRIVE for scalable training.

Disentangling Loss from Forgetting. While correlated ($r \approx 0.65$), loss magnitude alone does not fully explain forgetting; we posit gradient conflict as the primary driver. To verify this, we isolated the top 40% high-loss samples from Flowers102 and partitioned them by Gradient Inner Product (GIP) with the retention task. This yielded a Conflicting Group ($GIP < 0$, 293 images) that opposes retention, and a Synergistic Group ($GIP > 0$, 117 images) geometrically aligned with it.

As summarized in Table 25, we evaluate performance by training on distributions combining the remaining 60% low-loss samples with either the Synergistic or Conflicting high-loss subgroups. Fine-tuning on the Synergistic mixture achieves competitive target accuracy while maintaining ImageNet performance comparable to the pre-trained baseline.

In contrast, the Conflicting mixture triggers severe forgetting. This dichotomy confirms that high-loss samples can be learned with limited forgetting when their gradients align with the retention task.

Table 25: Impact of Gradient Alignment. Synergistic samples prevent forgetting despite high loss.

Subset	Target (%)	Retain (%)
Pre-trained	-	80.10
Conflicting	84.85	38.42
Synergistic	83.76	79.15

Conflict Analysis via Min-Max Normalization. To intuitively visualize the trade-off landscape independent of raw gradient magnitudes, we project the accumulated Shapley values into the unit interval $[0, 1]$ using Min-Max normalization. Let $\mathbf{a}^{\text{shap}}$ and $\mathbf{t}^{\text{shap}}$ be the vectors containing the contributions of all samples. The normalized scores are defined as:

$$\widetilde{a}_i := \frac{a_i^{\text{shap}} - \min(\mathbf{a}^{\text{shap}})}{\max(\mathbf{a}^{\text{shap}}) - \min(\mathbf{a}^{\text{shap}})}, \qquad \widetilde{t}_i := \frac{t_i^{\text{shap}} - \min(\mathbf{t}^{\text{shap}})}{\max(\mathbf{t}^{\text{shap}}) - \min(\mathbf{t}^{\text{shap}})}. \tag{29}$$

Fig. 7 plots the normalized task-adaptation contribution $\widetilde{t}_i$ versus the anti-forgetting contribution $\widetilde{a}_i$. The regression analysis reveals a pronounced negative correlation (Pearson's $r = -0.628$, $R^2 = 0.394$), supporting a strong empirical conflict between the two objectives: samples conducive to adaptation often tend to be less favorable for retention. The data distribution is heavily skewed toward the off-diagonal quadrants: 44.8% of samples cluster in the Adaptation-Biased region (top-left) and 31.4% in the Retention-Biased region (bottom-right). Conversely, the Co-contributing region (top-right) remains sparse, containing only 8.7% of the data. This scarcity of synergistic samples underscores the stability-plasticity dilemma and motivates DRIVE's dynamic trade-off mechanism, which navigates the conflicting landscape more effectively than naive uniform averaging.

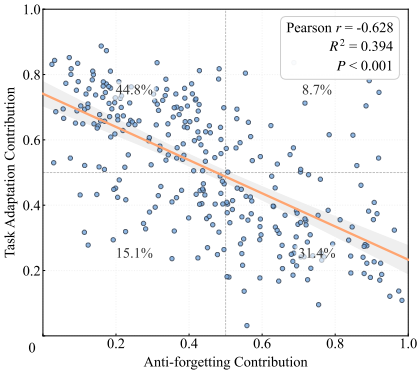


Figure 7: Scatter plot of normalized Anti-forgetting ($\widetilde{a}_i$) vs. Task Adaptation ($\widetilde{t}_i$) contributions.

F.2 Sample Diversity

Sample diversity is a multi-faceted construct that ensures the model explores the parameter space effectively while avoiding overfitting to redundant patterns [62, 68]. In this section, we expand on the metrics used to quantify diversity in recent literature.

F.2.1 Diversity in Fine-tuning

1. Semantic and Clustering-based Metrics. These metrics rely on the distance between samples in an embedding space to ensure coverage [22, 23]. [23] identify representative core sets via clustering-based strategies to improve sample efficiency [23]. While effective, these methods are often discrete and non-differentiable. DEFT-UCS utilizes core-set selection for text-editing tasks to reduce data size while preserving performance [22]. However, it lacks the continuous, joint optimization framework provided by DRIVE.

2. Gradient-space Diversity. Diversity in the gradient space is often a more direct proxy for optimization robustness [50, 109]. [109] demonstrate that selecting data based on gradient alignment can target specific downstream capabilities. DRIVE captures a similar signal through the task-adaptation term t_i, which prioritizes samples essential for mastering the new task. GSS prioritizes samples whose gradients maximally align with the feasible regions of previous tasks to mitigate forgetting [50].

3. Multi-level Diversity (Macro to Micro). Recent trends categorize diversity across hierarchies of the instruction dataset [67]. Studies show that increasing response diversity (micro-diversity) is more beneficial for human alignment than simply expanding prompt quantity (macro-diversity) [68]. Methods like DEITA prioritize semantic complexity and quality over volume, effectively reducing the sample size needed for fine-tuning [21, 24].

While existing methods often rely on selection-based or discrete metrics, DRIVE integrates diversity directly into the objective function via a differentiable covariance determinant, providing a unified and scalable solution.

F.2.2 Details of DRIVE-Cluster

In the replacement study, we compare DRIVE with DRIVE-Cluster, a heuristic baseline that enforces diversity through rigid geometric constraints. This variant corresponds to the DRIVE framework with $\eta = 0$, where the diversity requirement is encoded directly into the prior distribution rather than a dynamic regularizer.

Formulation. Given K non-empty clusters $\{C_k\}_{k=1}^{K}$ obtained via K-Means clustering, we construct a cluster-balanced prior $\boldsymbol{p} = \{p_i\}_{i=1}^{N}$ over the entire dataset. To ensure each semantic cluster holds equal probability mass $(1/K)$ initially, the prior is defined as:

$$p_i = \frac{1}{K|C_k|}, \quad \forall i \in C_k. \tag{30}$$

The weight optimization is then performed on the entire dataset by minimizing the KL-divergence against this global prior, subject to a single global normalization constraint:

$$\max_{\boldsymbol{w}} \quad \sum_{i=1}^{N} w_i c_i - \tau KL(\boldsymbol{w}||\boldsymbol{p}) \quad \text{s.t.} \quad \sum_{i=1}^{N} w_i = 1, \quad w_i \geq 0, \quad \forall i. \tag{31}$$

This formulation unifies all samples into a single optimization landscape, allowing them to compete globally based on their contribution scores, modulated by the cluster-balanced prior.

Closed-form Solution. By introducing a Lagrange multiplier for the global constraint $\sum w_i = 1$, we derive the optimal solution. The stationarity condition yields:

$$w_i^* \propto p_i \exp(c_i/\tau). \tag{32}$$

Enforcing the global normalization condition, we obtain the standard Softmax form:

$$w_i^* = \frac{p_i \exp(c_i/\tau)}{\sum_{j=1}^{N} p_j \exp(c_j/\tau)}. \tag{33}$$

This result highlights that DRIVE-Cluster effectively weights samples by scaling their exponential contribution with the prior p_i. Since $p_i \propto 1/|C_k|$, samples from dense (redundant) clusters are naturally down-weighted in the global competition, ensuring that the final weight distribution maintains semantic diversity without requiring explicit cluster-level hard constraints.

F.3 Analysis of Covariance Approximation Strategies

In Section 3.3, we employ randomized dimensionality reduction via the Johnson-Lindenstrauss (JL) transform to approximate the high-dimensional feature space $\mathbb{R}^d$ in a lower-dimensional subspace $\mathbb{R}^k$. This design choice balances computational efficiency with geometric fidelity. In this section, we justify this choice by comparing it against two alternative strategies: a computationally cheaper diagonal approximation and a theoretically more robust ensemble projection method.

F.3.1 Comparison with Diagonal Approximation

A computationally cheaper alternative to our randomized projection strategy is the diagonal approximation, which assumes feature independence by setting off-diagonal elements of the weighted covariance matrix $\mathbf{\Sigma}_w$ to zero. This approach is widely adopted in methods like EWC [38] and scalable Laplace approximations [110] to reduce the storage and inversion cost from $O(d^3)$ to $O(d)$:

$$\log|\mathrm{diag}(\boldsymbol{\sigma}^2)| = \sum_{j=1}^{d} \log(\sigma_j^2). \tag{34}$$

However, this approximation creates a critical failure mode for diversity weighting: it cannot distinguish between diverse orthogonal features and redundant collinear features, provided their marginal variances are identical. Since DRIVE aims to explicitly penalize geometric redundancy (i.e., feature correlation), preserving off-diagonal interactions via random projection is theoretically necessary to enforce D-Optimality. Consequently, despite its linear complexity, the diagonal approximation is unsuitable for our objective.

F.3.2 Empirical Analysis of Single vs. Ensemble Random Projections

While a single random projection preserves pairwise distances with high probability, ensembling multiple independent projections could theoretically reduce the approximation variance. We investigate whether this potential gain justifies the added computational cost.

Experimental Setup. To evaluate the impact of ensembling, we compare the performance of DRIVE using a single Gaussian random projection matrix $\mathbf{R} \in \mathbb{R}^{d\times k}$ against an ensemble of M independent matrices $\{\mathbf{R}_m\}_{m=1}^{M}$. For the ensemble method, the diversity regularization term averages the log-determinants across M projections:

$$\mathcal{J}(w) = \frac{1}{M}\sum_{m=1}^{M} \eta \log\left|\sum_{i=1}^{N} w_i(\mathbf{R}_m^\top \bar{\boldsymbol{x}}_i)(\mathbf{R}_m^\top \bar{\boldsymbol{x}}_i)^\top\right| \tag{35}$$

We conducted experiments on ResNet-50 fine-tuned on CIFAR-100 and Flowers102, setting $k = 128$ and varying $M \in \{1, 5, 10\}$. Other hyperparameters remain consistent with the main experiments.

Table 26: Performance of DRIVE with varying ensemble sizes on CIFAR-100 and Flowers102.

Method	CIFAR-100			Flowers102		
	Target	IN-1K	Avg.	Target	IN-1K	Avg.
Single Proj. ($M = 1$)	83.03	61.64	72.34	92.67	75.02	83.85
Ensemble ($M = 5$)	83.08	**61.67**	**72.38**	92.65	**75.09**	83.87
Ensemble ($M = 10$)	**83.10**	61.65	**72.38**	**92.71**	75.05	**83.88**

As observed in Table 26, increasing the ensemble size from $M = 1$ to $M = 10$ yields negligible performance gains. Specifically, on CIFAR-100, the average accuracy improves by only 0.04%, and on Flowers102, the improvement is marginal with a gain of just 0.03%. This performance plateau suggests that a single Gaussian projection ($M = 1$) is sufficiently robust for capturing the geometric structure required by our diversity regularizer. The variance introduced by a single projection does not destabilize the optimization trajectory. Therefore, given the lack of significant benefits from ensembling, we use a single random projection in our main experiments to maintain model simplicity.

F.4 Empirical Verification of Dynamic Adaptation on Hard Samples

To empirically validate FLOW's under-fitting behavior on challenging samples, we conducted an experiment monitoring the optimization dynamics on the most difficult subset of the data.

Setup. We fine-tune a ResNet-50 backbone on CIFAR-100. Before fine-tuning, we select the top 10% training samples (approximately 5,000 images) that incur the largest loss under the pre-trained model $\boldsymbol{\theta}_{\text{pre}}$, and define this subset as a fixed hard-sample probe set $\mathcal{D}_{\text{hard}}$ (not used for model selection or hyperparameter tuning). During training, for both FLOW and DRIVE, we monitor the raw (unweighted) mean loss on $\mathcal{D}_{\text{hard}}$ over time.

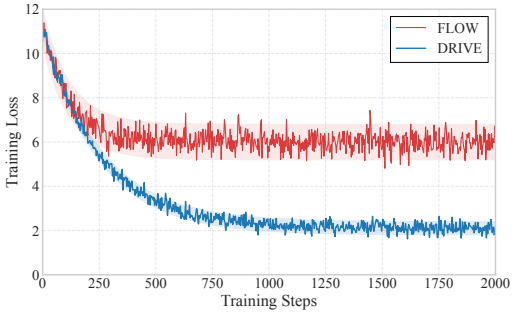


Figure 8: Loss Dynamics on $\mathcal{D}_{hard}$. FLOW exhibits high-variance stalling (jagged), while DRIVE achieves stable adaptation (smooth).

As illustrated in Figure 8, the distinct trajectories confirm our theoretical analysis. The loss curve for FLOW remains nearly flat, stagnating at a high value (≈ 6.0). Notably, the curve exhibits severe fluctuations and instability. This indicates that while the model attempts to update, the ill-conditioned landscape causes the optimization to oscillate without making effective progress. By assigning negligible weights ($w_i \approx 0$) to high-loss samples, FLOW effectively freezes the model parameters in the "conflicting subspace," preventing the gradient updates required to master these samples.

In contrast, DRIVE exhibits a steady, smooth, and significant decrease, with the final loss dropping to approximately 2.0. This substantial reduction serves as direct evidence that the model has successfully "learned" the underlying patterns of these hard samples.

To provide a mechanistic explanation for the optimization stalling observed in FLOW, we monitored the condition number $\kappa(\Sigma_w) = \lambda_{\max}/\lambda_{\min}$ of the weighted covariance matrix. Our empirical observations reveal a stark contrast in the optimization geometry: FLOW exhibits spectral collapse, resulting in an ill-conditioned landscape with $\kappa(\Sigma_w)$ reaching the order of 10^6. This anisotropy traps the optimization in a steep valley, causing the gradients to bounce ineffectively between valley walls (zig-zagging behavior). In contrast, thanks to the determinant-based diversity regularization, DRIVE maintains a better-conditioned geometry, stabilizing $\kappa(\Sigma_w)$ at the order of 10^3. This reduction by three orders of magnitude ($10^6 \rightarrow 10^3$) empirically validates that DRIVE enforces D-Optimality, thereby flattening the loss landscape and ensuring robust convergence.

F.5 Sensitivity to the Projection Dimension k

We conduct a lightweight sensitivity study on Gemma-2-2B fine-tuned on MetaMathQA, varying the sketch dimension k while keeping all other settings fixed. As shown in Table 27, overly small sketches ($k = 32, 64$) lead to clear drops in both target accuracy and overall performance, indicating that the weighted local covariance geometry is compressed too aggressively. In contrast, increasing the sketch size to $k = 128$ yields a substantial performance jump, suggesting that this scale is sufficient to preserve the key local geometry used by DRIVE. Further increasing k to 256 or 512 brings only marginal gains, while the training cost rises sharply. Therefore, $k = 128$ emerges as the practical elbow point, offering the best trade-off between approximation fidelity and efficiency, and thus serves as a well-justified default choice in DRIVE.

Table 27: Sensitivity to k on Gemma-2-2B / MetaMathQA. Relative time is normalized by $k = 32$.

k	Gen.	Target	Overall	Time
32	46.72	54.18	50.45	1.00
64	47.01	57.60	52.31	1.30
128	47.64	63.46	55.55	1.70
256	47.69	63.79	55.74	2.50
512	47.71	64.32	56.02	4.10

G Limitations

While DRIVE demonstrates consistent gains across vision models and LLMs, we note several limitations. First, our formal guarantees are established in linear and locally linearized nonlinear regimes. The local nonlinear analysis extends the purely linear case by retaining explicit Hessian remainder control, but it still relies on smoothness and local-approximation assumptions, and is valid only

within a bounded neighborhood of the pre-trained weights. Extending these results to the fully non-convex global dynamics of large transformers remains an open challenge; our support beyond the local regime is currently empirical.

Second, to ensure tractability, we employ the Johnson–Lindenstrauss projection and estimate diversity with a finite FIFO buffer. Although the projection preserves dominant low-rank pairwise structures, it may attenuate diversity signals when the effective intrinsic dimension is high. The FIFO buffer further captures local rather than global diversity and may miss distant redundant samples in the data stream. Moreover, the method's effectiveness relies on reasonable alignment between the pre-trained representation and the target domain.

Finally, diversity introduces computational overhead from weighted covariance and determinant estimation. While dimensionality reduction keeps this cost moderate (e.g., $\sim$17% for LLMs) in our reported settings, further acceleration would be beneficial for strictly constrained training budgets.

Broader Impacts

DRIVE improves the efficiency and reliability of data-oblivious fine-tuning by assigning adaptive sample weights, which can reduce unnecessary training effort and make model adaptation more accessible. At the same time, more efficient fine-tuning may also lower the barrier to harmful, biased, or low-quality downstream adaptation. Our experiments use public datasets and standard benchmarks, and do not collect user data or involve human-subject studies.

Reproducibility

We will release anonymized code, configuration files, training scripts, and evaluation scripts for reproducing the main tables and diagnostic experiments. All reported experiments use public datasets and publicly available base models, with implementation details provided in the main text and appendices.

Licenses and Assets

We use publicly available datasets, benchmarks, and base models, following their respective licenses and terms of use. The proposed method does not introduce a new dataset or release new model checkpoints. The released assets consist of anonymized code, configurations, and scripts for reproducing the experiments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3, Section 4 and Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 4, Section B, Section C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5, Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We used the public dataset, and we provide our code at https://anonymous.4open.science/r/DRIVE-158.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer:[Yes]

   Justification: Section 5, Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: We report the main experimental results with error bars, where each value is reported as mean ± standard deviation over five independent runs. The reported variability captures randomness from independent runs under the same experimental setting, including initialization, data order, and stochastic optimization.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: There are no experiments in this work requiring human participants. All the datasets are in the public domain and have been used in accordance with their respective licenses.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention that our work could have a positive impact by improving the efficiency and robustness of data-oblivious fine-tuning, especially through adaptive data weighting that helps mitigate catastrophic forgetting. These benefits may make model customization more reliable when pre-training data are unavailable and contribute to more manageable safety and security considerations during downstream adaptation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the paper, we specify the sources of the open-source data used and provide proper citations. Regarding the code, we will acknowledge the relevant contributors on our GitHub repository when we make it publicly available.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper releases the implementation of DRIVE through an anonymized code repository. The repository provides documentation for the proposed method, running instructions, required dependencies, and experimental reproduction.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human subjects, user studies, surveys, crowdsourcing, or private identifiable data. Experiments use existing public datasets and pretrained models, so IRB approval is not applicable.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used only for grammar correction and refinement.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.